wellcome trust sanger institute

Workshop

on Bacterial Genomics

28-30 September 2005

Held at:

Ciutat de les Arts i de les Ciències - Valencia (Spain)

Timetable

Wednesday, 28th September

08:30-09:30 Registration
09:30-10:00 Open ceremony. Andrés Moya.
10:00-10:30 Coffee
10:30-11:00 Introduction. Julian Parkhill
11:00-14:00 Artemis: Guided excercises. Nicholas Thomson.
14:00-15:00 Lunch
15:00-16:00 Gene Prediction. Nicholas Thomson.
16:00-17:00 ACT: Guided exercises. Ana Cerdeño.
17:00-17:30 Coffee
17:30-19:30 ACT: Guided exercises (cont'd). Ana Cerdeño.

19:30-20:30. **Public conference. Julian Parkhill**

21:00 Official dinner. L'Oceanogràfic. Submarine Restaurant.

Thursday, 29th September

09:00-10:30 ACT: Guided exercises (cont'd). Ana Cerdeño
10:30-11:00 Coffee
11:00-12:30 Generating ACT comparison files. Ana Cerdeño.
12:30-14:00 Jemboss + Internet Genome Resources. Tim Carver.
14:00-15:00 Lunch
15:00-16:30 Internet Genome Resources. Ana Cerdeño.
16:30-17:00 Coffee
17:00-19:00 Data mining using GeneDB. Christopher Peacock.

19:30-20:30. **Public conference. Jean Marie Claverie**.

Visit to two Scientific Centers in Valencia
19:30-22:30. Visit to CSAT and IVI.

Science Bar
21:00-22-15. Museum's Bar.

Friday, 30th September

09:00-10:30 Comparative genomics. Francisco Silva and Amparo Latorre
10:30-11:00 Coffee
11:00-12:30 Comparative genomics (cont'd). Francisco Silva and Amparo Latorre.
12:30-14:00 Phylogenomics. Fernando González and Rosario Gil.
14:00-15:00 Lunch
15:00-16:30 Phylogenomics (cont'd). Fernando González and Rosario Gil
16:30-17:30 Genome flexibility. Alex Mira.
17:30-19:30 Annotation summary exercise or own seq + Mop-up.

Glossary of Abbreviations and Terms

ACT	Artemis Comparison Tool.
BLAST	Basic local alignment search tool.
CDS	Coding sequence (Gene with no biological evidence for expression).
CNRS	Centre National De La Recherche Scientifique.
DDBJ	DNA data bank of Japan.
EBI	European Bioinformatics Institute, Hinxton. An outstation of the European Molecular Biology Laboratory.
EMBL	European Molecular Biology Laboratory, the name of the European DNA database.
EST	Expressed sequence tag.
Fasta	Part of the 'Fast' repertoire of global alignment search tools.
Flatfile	A simple text file used as an alternative to a database to storing data.
GENE-IT	Is a company that collaborates with the EBI and others to discover the functions of genes through comparative genomics.
HMM	Hidden Markov Model.
INRA	Institut National De La Recherche Agronomique.
InterPro	A search tool which brings together many of the commonly used signature databases for sequence searching.
LINUX	A packaged version of UNIX for the PC.
mRNA	messenger RNA, processed RNA molecule to be translated to form protein.
NCBI	National Centre for Biotechnology Information. Part of the U.S. National Library of Medicine (NLM), National Institutes of Health (NIH).
PRINTS	Proteins Finger print database, a compendium of protein finger prints.
PFAM	Protein family, a searchable database of protein domains.
ProDom	a comprehensive set of protein domain families.
Prosite	Database of protein families and domains.
PSU	Pathogen Sequencing Unit.
RFAM	A searchable database of RNA families.
SIB	Swiss Institute of Bioinformatics (SIB).
SignalP	A program to predict the presence and location of signal peptide cleavage sites in amino acid sequences.
SMART	Simple Modular Architecture Tool.
SWISS-PROT	is a curated protein sequence database.
TIGR	The Institute of Genome Research.
TIGRfam	The Institute of Genome Research protein family database.
TMHMM	Program for prediction of transmembrane helices in proteins.
TrEMBL	Computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.
UNIX	A computer operating system.

Index

Index

Module 1: *Artemis*
Exercise 1
Exercise 2
Module 2: *Comparative Genomics*
Exercise 1
Exercise 2
Exercise 3
Exercise 4
Module 3: *Generating ACT comparison files using BLAST*
Exercise 1
Exercise 2
Exercise 3
Exercise 4
Module 4: *Jemboss*
Exercise 1
Exercise 2
Module 5: *Genome Resources*
Section 1
Exercise 1
Exercise 2
Exercise 3
Section 2
Exercise 1
Exercise 2
Section 3
Module 6: *Data Mining using GeneDB*

References
Appendices

Module 1
Artemis

Introduction

Artemis (Rutherford *et al*., (2000) is a DNA viewer program, written by Kim Rutherford, and used for both Prokaryotic and Eukaryotic annotations. It allows the user to get away from the relatively faceless EMBL and Genbank style database files and view the sequence in a graphical and highly interactive format. Artemis is designed to present multiple lines of information within a single context. This manifests itself as being able to zoom in to look for fine DNA motifs as well as being able to zoom out and bring into view operons, several kilobases of a genome or in fact to view an entire genome in one screen. It is also possible to perform quite sophisticated analyses and store the output within the 'Artemis environment' to be accessed later.

Aims

The aim of this Module is for you to become familiar with the basic functioning of Artemis by using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and crannies of Artemis that are not featured in the exercises in this manual. Like all the Modules in this workshop, the key is ‘if you don’t understand please ask’.

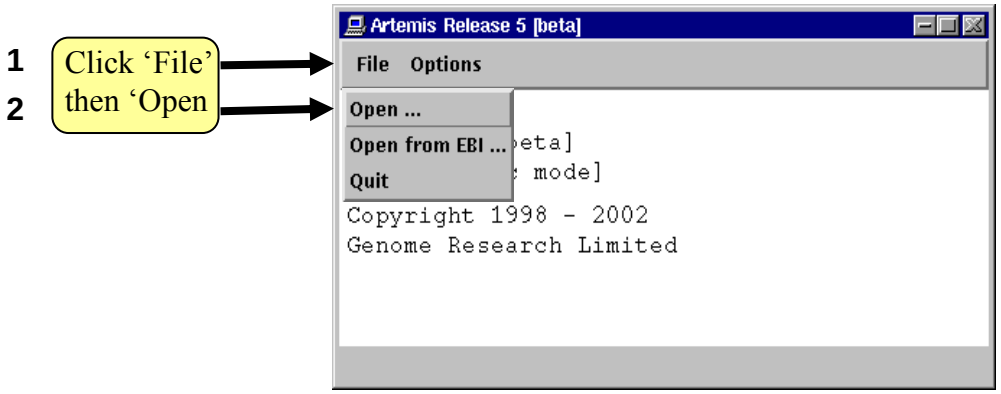
Artemis Exercise 1 Part I

1. Starting up the Artemis software

Navigate your way into the correct directory for this module
Then type:

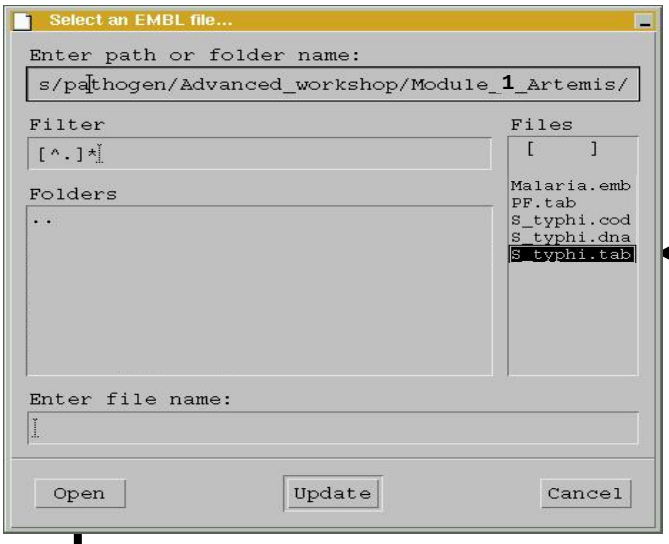
art & [return]

A small start-up window will appear (see below).
Now follow the sequence of numbers to load up the *Salmonella typhi* chromosome sequence.
Ask a demonstrator for help if you have any problems.



For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.

In the 'Options' menu you can switch between prokaryotic and eukaryotic mode.



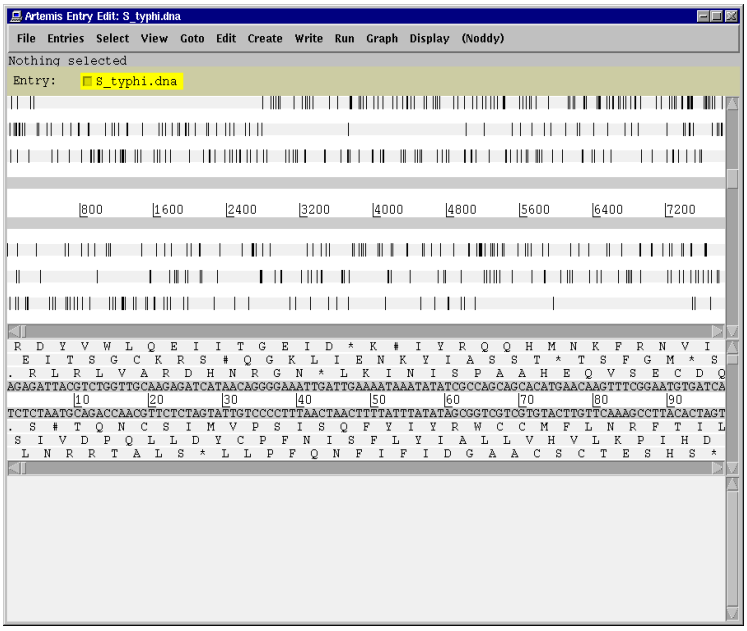
3 Single click to select DNA file

DNA sequence files will have the suffix '.dna'. Annotation files end with '.tab'.

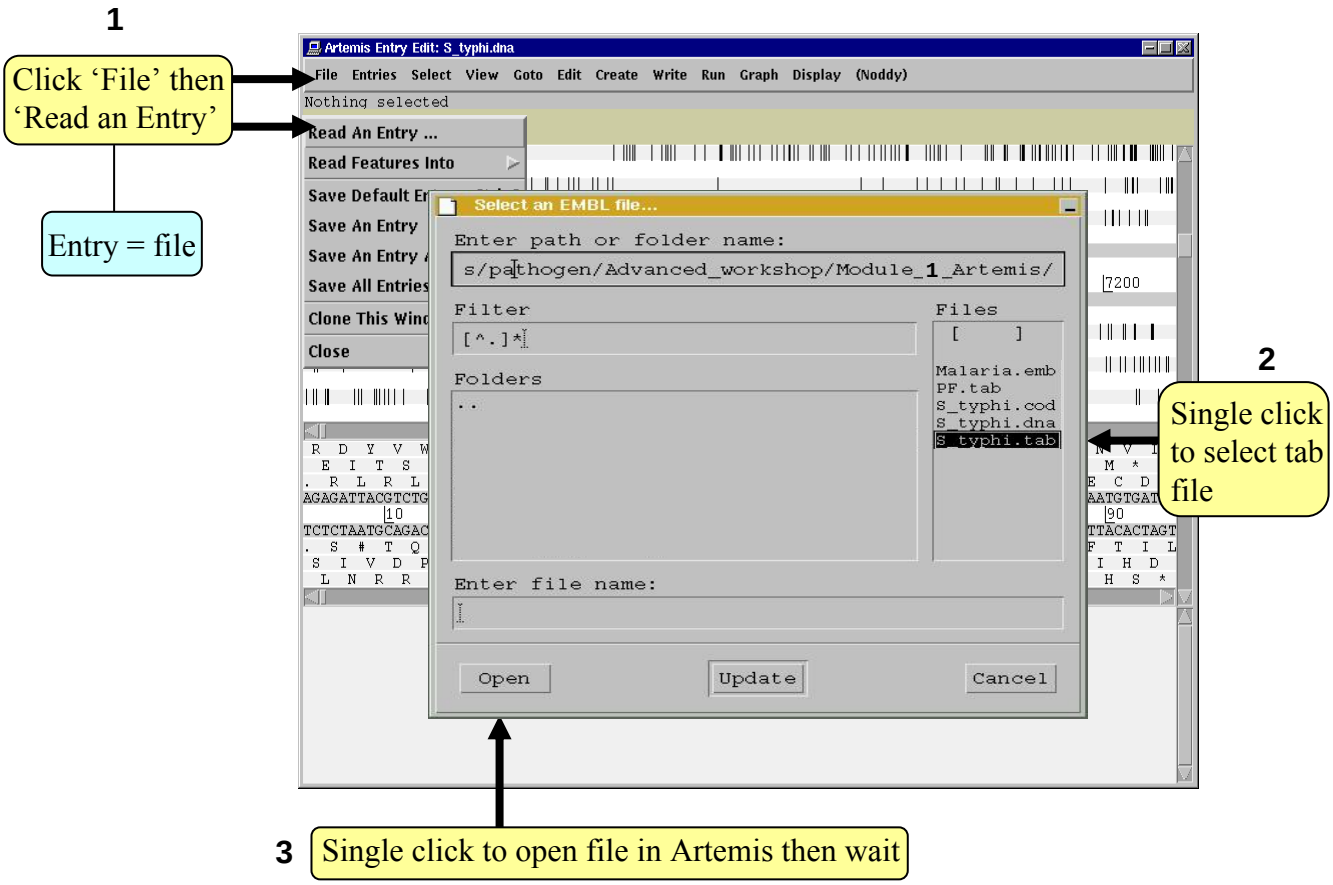
4 Single click to open file in Artemis then wait

2. Loading annotation files (entries) into Artemis

Hopefully you will now have an Artemis window like this! If not, ask a demonstrator for assistance.



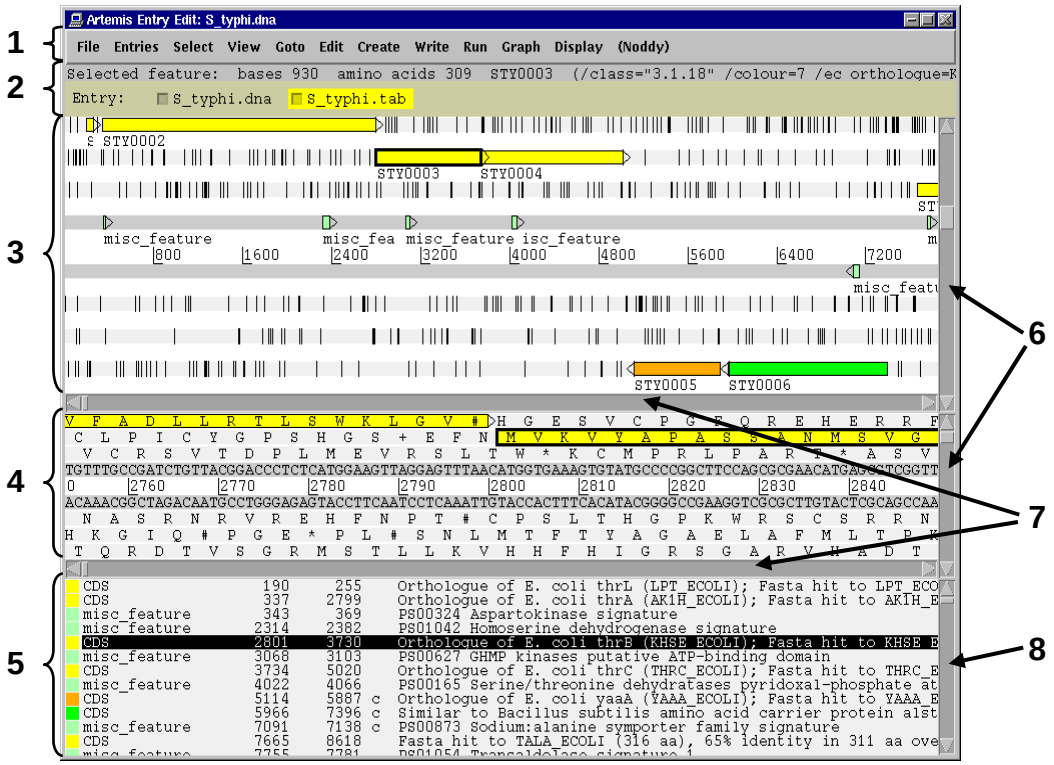
Now follow the numbers to load up the annotation file for the *Salmonella typhi* chromosome.



What's an "Entry"? It's a file of DNA and/or amino acid features which can be overlaid onto the sequence information displayed in the main Artemis view panel.

3. The basics of Artemis

Now you have an Artemis window open let's look at what's in there.



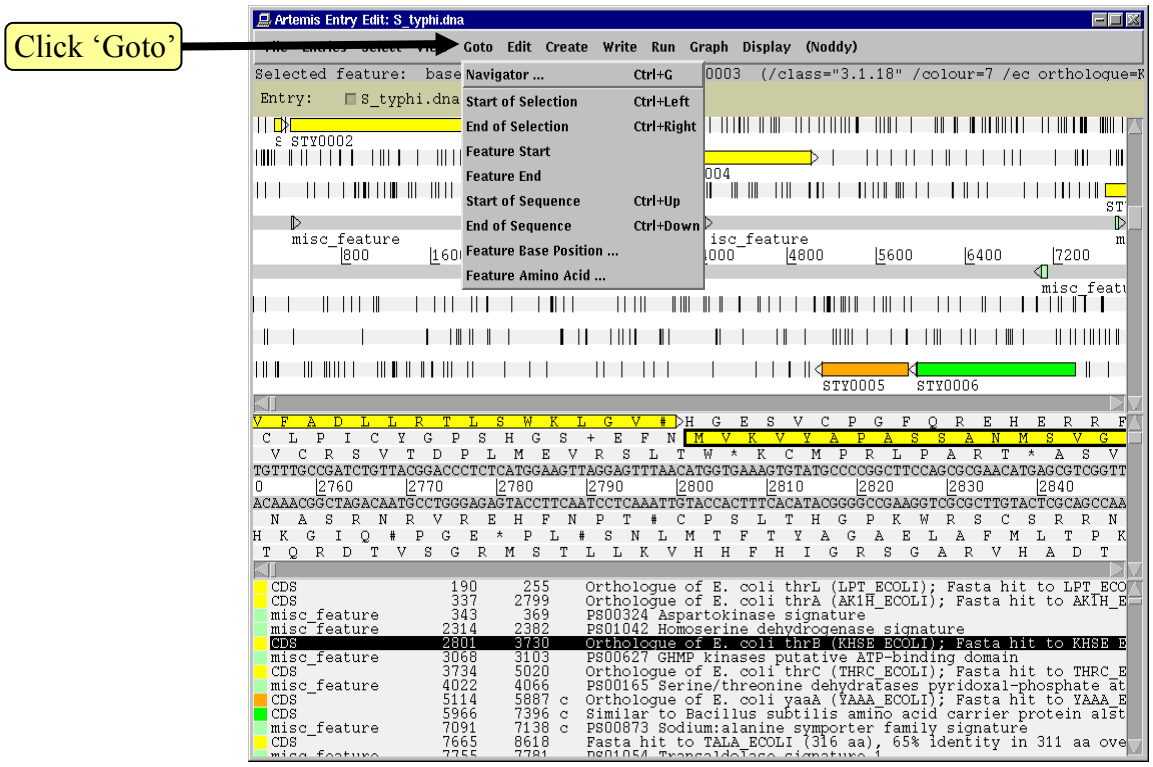
1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case gene STY0003 (top line).
3. This is the main sequence view panel. The central 2 grey lines represent the forward (top) and reverse (bottom) DNA strands. Above and below those are the 3 forward and 3 reverse reading frames. Stop codons are marked as black vertical bars. Genes and other features (eg. Pfam and Prosite matches) are displayed as coloured boxes. We will refer to genes as coding sequences or CDSs from now on.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene highlighted. The list can be scrolled (8, below).
6. Sliders for zooming view panels.
7. Sliders for scrolling along the DNA.
8. Slider for scrolling feature list.

4. Getting around in Artemis

The 3 main ways of getting to where you want to be in Artemis are the Goto dropdown menu, the Navigator and the Feature Selector. The best method depends on what you're trying to do and knowing which one to use comes with practice.

4.1 The 'Goto'menu

The functions on this menu (ignore the Navigator for now) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. This one's really intuitive so give it a try!



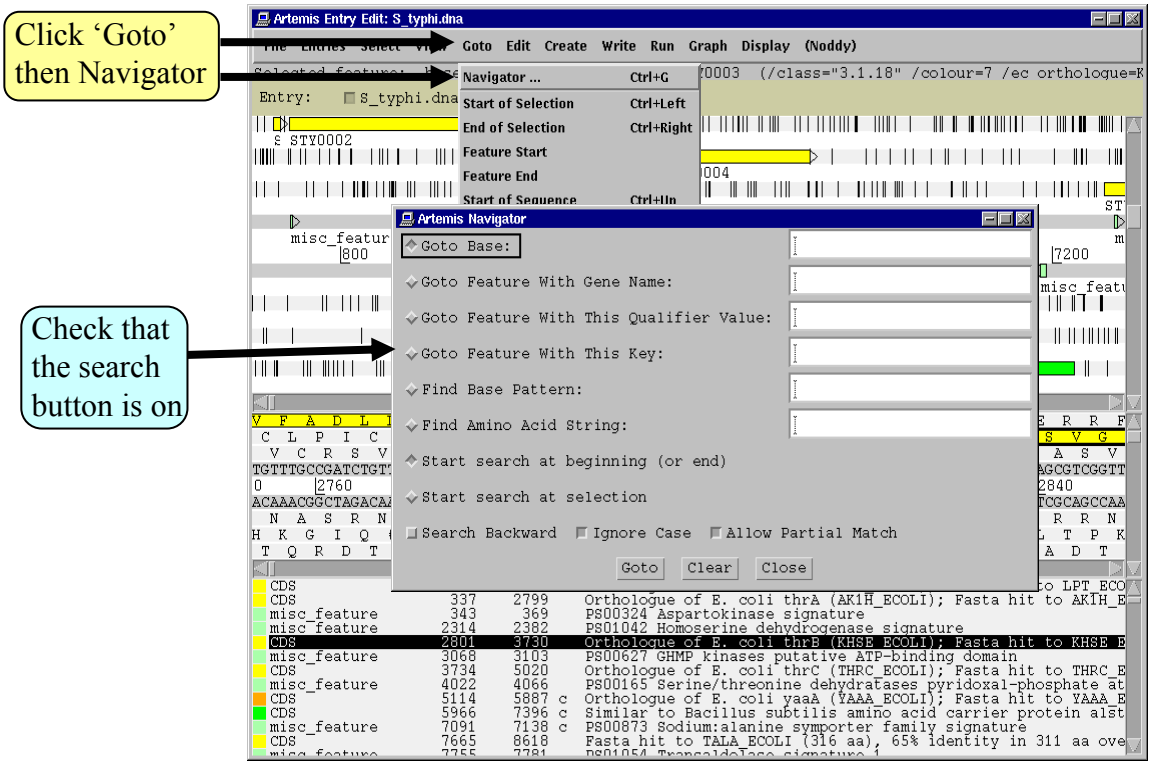
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try!

Suggested tasks:

1. Zoom out, highlight a large region of sequence by clicking the left hand button and dragging the cursor then go to the start and end of the highlighted region.
2. Select a gene then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a gene. Within it, go to a base (nucleotide) and/or amino acid of your choice.

4.2 Navigator

The Navigator panel is fairly intuitive so open it up and give it a try.



Suggestions of where to go:

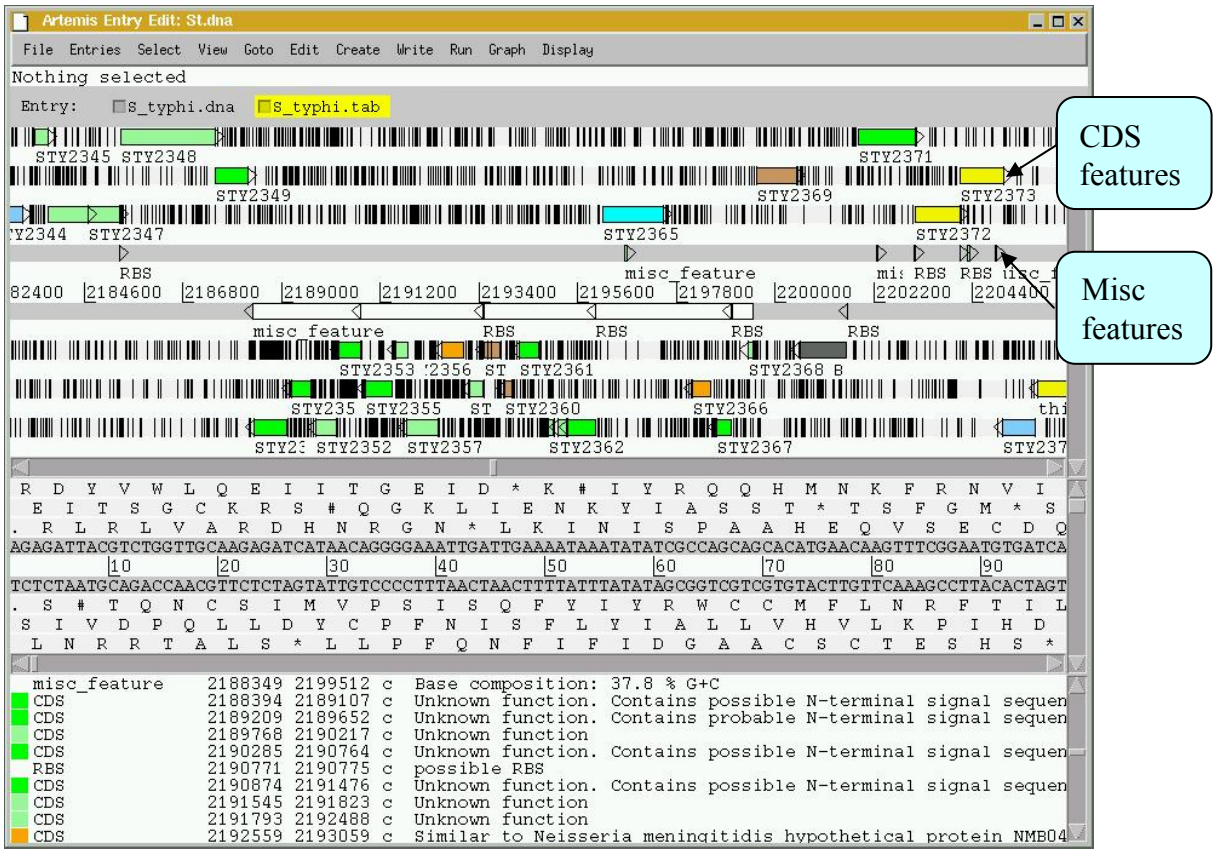
1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try 'fts').
3. Use 'Goto Feature With This Qualifier value' to search the contents of all qualifiers for a particular term. For example using the word 'pseudogene' will take you to the next feature with the word 'pseudogene' in any of its qualifiers. Note how repeated clicking of the 'Goto' button takes you through the pseudogenes as they occur on the chromsome.
4. tRNA genes. Type 'tRNA' in the 'Goto Feature With This Key'.
5. Regulator-binding DNA consensus sequence (real or made up!). Note that degenerate base values can be used (Appendix VIII).
6. Amino acid consensus sequences (real or made up!). You can use 'X's. Note that it searches all six reading frames regardless of whether the amino acids are encoded or not.

What are Keys and Qualifiers? See Appendix III

Clearly there are many more features of Artemis which we will not have time to explain in detail. Before getting on with this next section it might be worth browsing the menus. Hopefully you will find most of them easy to understand.

Artemis Exercise 1 Part II

This part of the exercise uses the files and data you already have loaded into Artemis from Part I. By a method of your choice go to the region located between bases 2188349 to 2199512 on the DNA sequence. This region is bordered by the *fbaB* gene which codes for fructose-bisphosphate aldolase. You can use either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.



Once you have found this region have a look at some of the information that is available to you:-

Information to view:

Annotation

If you click on a particular feature you can view the annotation attached to it: select a CDS feature (or any other feature) and click on the Edit menu and select Edit Selected Feature. A window will appear containing all the annotation that is associated with that CDS. The format for this information is constrained by that which can be submitted to the EMBL database as seen in Module 1.

Viewing amino acid or protein sequence

Click on the view menu and you will see various options for viewing the bases or amino acids of the feature you have selected, in two formats i.e. EMBL or FASTA. This can be very useful when using other programs that are not integrated into Artemis e.g. those available on the Web that require you to cut and paste sequence into them.

Plots/Graphs

Feature plots can be displayed by selecting a CDS feature then clicking 'View' and 'Show Feature Plots'. The window which appears shows plots predicting hydrophobicity, hydrophilicity and coiled-coil regions for the protein product of the selected CDS.

Load additional files

The results from Prosite searches run on the translation of each CDS should already be on display as pale-green boxes on the grey DNA lines. The results from the Pfam protein motif searches are not shown, but can be viewed by loading the appropriate file. Click on 'File' then 'Read an Entry' and select the file PF.tab. Each Pfam match will appear as a coloured blue feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'View Selection' or click 'Edit' then Edit Selected Features'. Please ask if you are unsure about Prosite and Pfam.

Viewing the results of database searches

Click the 'View' menu, then select 'Search Results' and then 'Fasta results'. The results of the database search will appear in a scrollable window. If you click on the button at the bottom of this window labelled 'view in browser', then the results will be posted into an internet browser window. Within this window there are many active links (coloured blue), to external sources of information such as the original database entries for all those aligning to your sequence, as well as information stored in PubMed, PFAM and many others. Have a play.

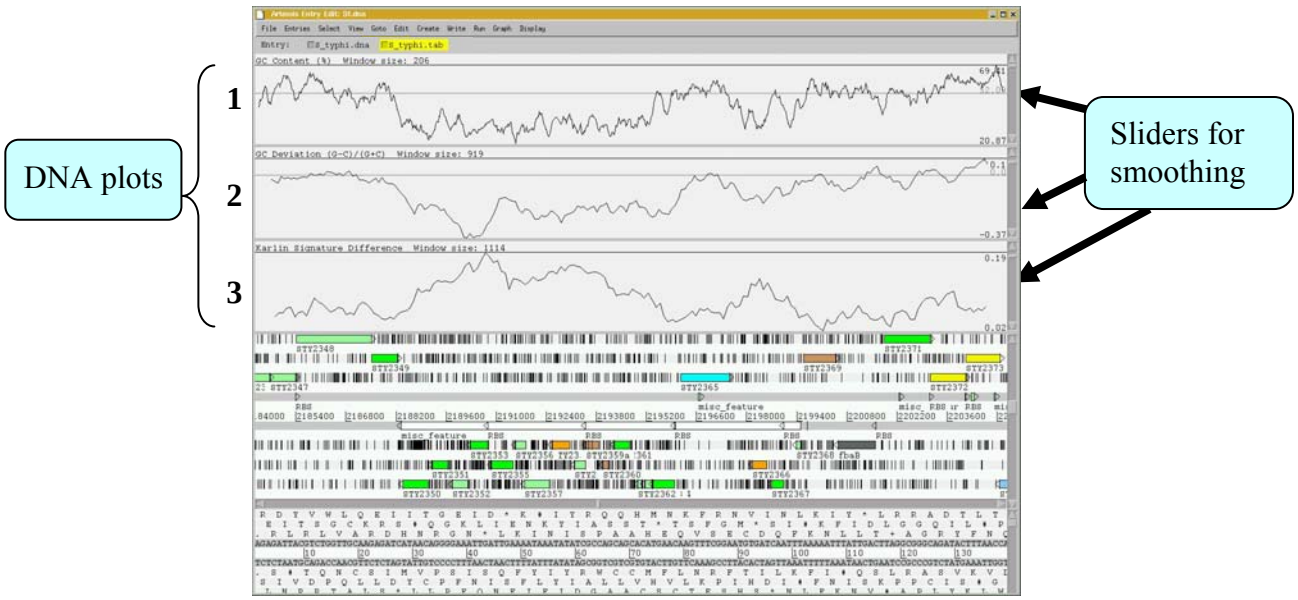
Further information on specific Prosite or Pfam entries can be found on the web at

http://ca.expasy.org/prosite and http://www.sanger.ac.uk/software/Pfam/tsearch.shtml

In addition to looking at the fine detail of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding in to the display various plots showing different characteristics of the DNA. This information is generated dynamically by Artemis and although this is a relatively speedy exercise for a small region of DNA, on a whole genome view (we will move onto this later) this many take a little time so be patient.

To view the graphs:

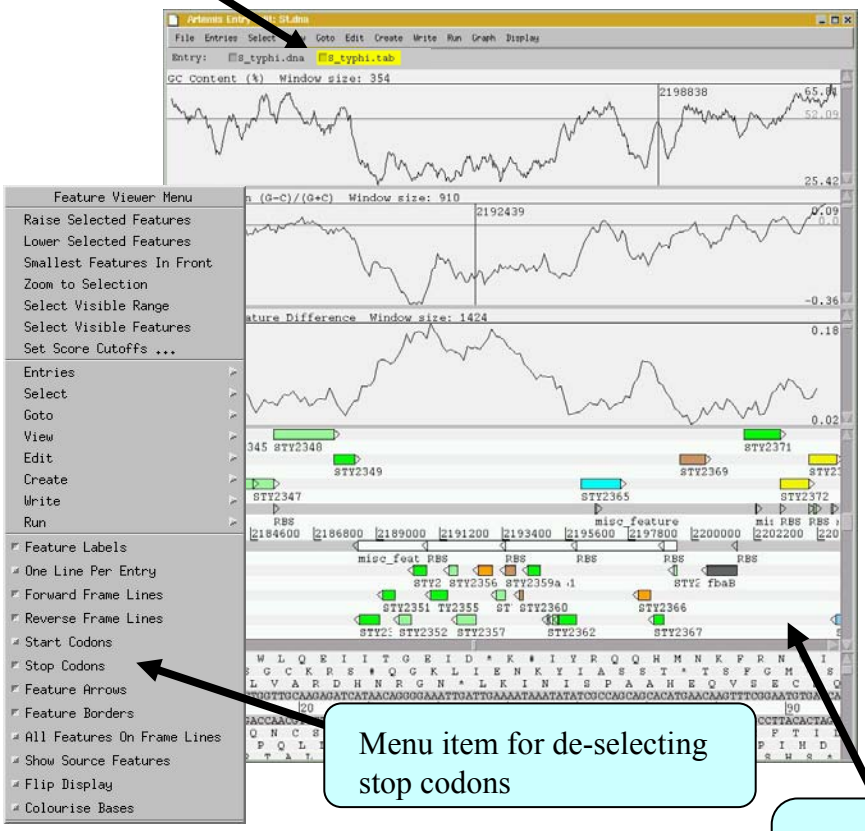
Click on the 'Graph' menu to see all those available. Perhaps some of the most useful plots are the 'GC Content (%)' (**1**) 'GC Deviation' (**2**) and 'Karlin signature plots' (**3**) as shown below. To adjust the smoothing of the graph you change the window size over which the points on the graph are calculated, using the sliders shown below. If you are not familiar with any of these please ask.



Notice how several of the plots show a marked deviation around the region you are currently looking at. To fully appreciate how anomalous this region is move the genome view by scrolling to the left and right of this region. The apparent unusual nucleotide content of this region is indicative of laterally acquired DNA that has inserted into the genome.

As well as looking at the characteristics of small regions of the genome, it is possible to zoom out and look at the characteristics of the genome as a whole. To view the entire genome use the sliders indicated below. However, be careful zooming out quickly with all the features being displayed, as this may temporarily lock up the computer. To make this process faster, and clearer, switch off stop codons by clicking with the right mouse button in the main view panel. A menu will appear with an option to de-select stop codons (see below). If you have any problems ask a demonstrator.

To de-select the annotation click here.

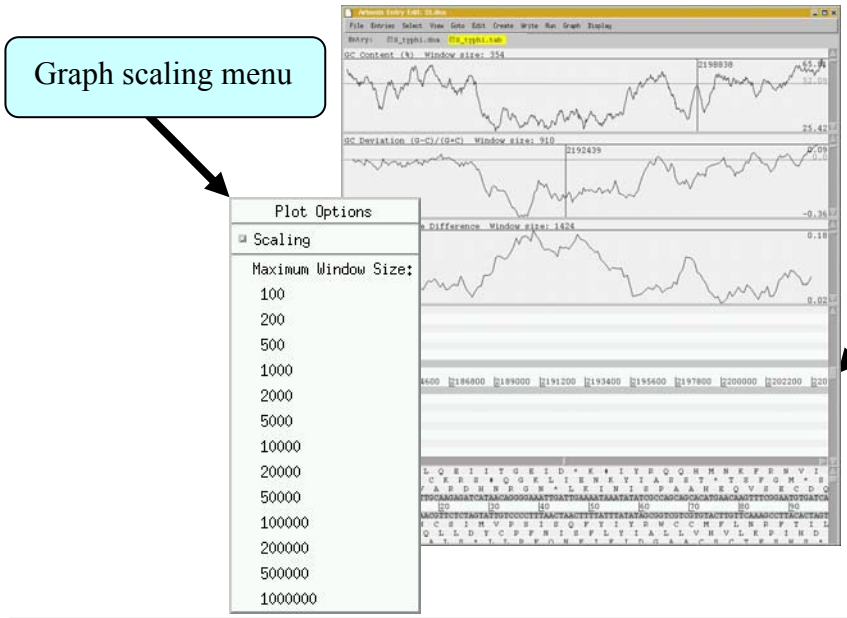


Menu item for de-selecting stop codons

No stop codons shown on frame lines

You will also need to temporarily remove all of the annotated features from the Artemis display window. In fact if you leave them on, which you can, they would be too small to see when you zoomed out to display the entire genome. To remove the annotation click on the S_typhi.tab entry button on the grey entry line of the Artemis window shown above.

Your Artemis window should now look similar to the one shown below.



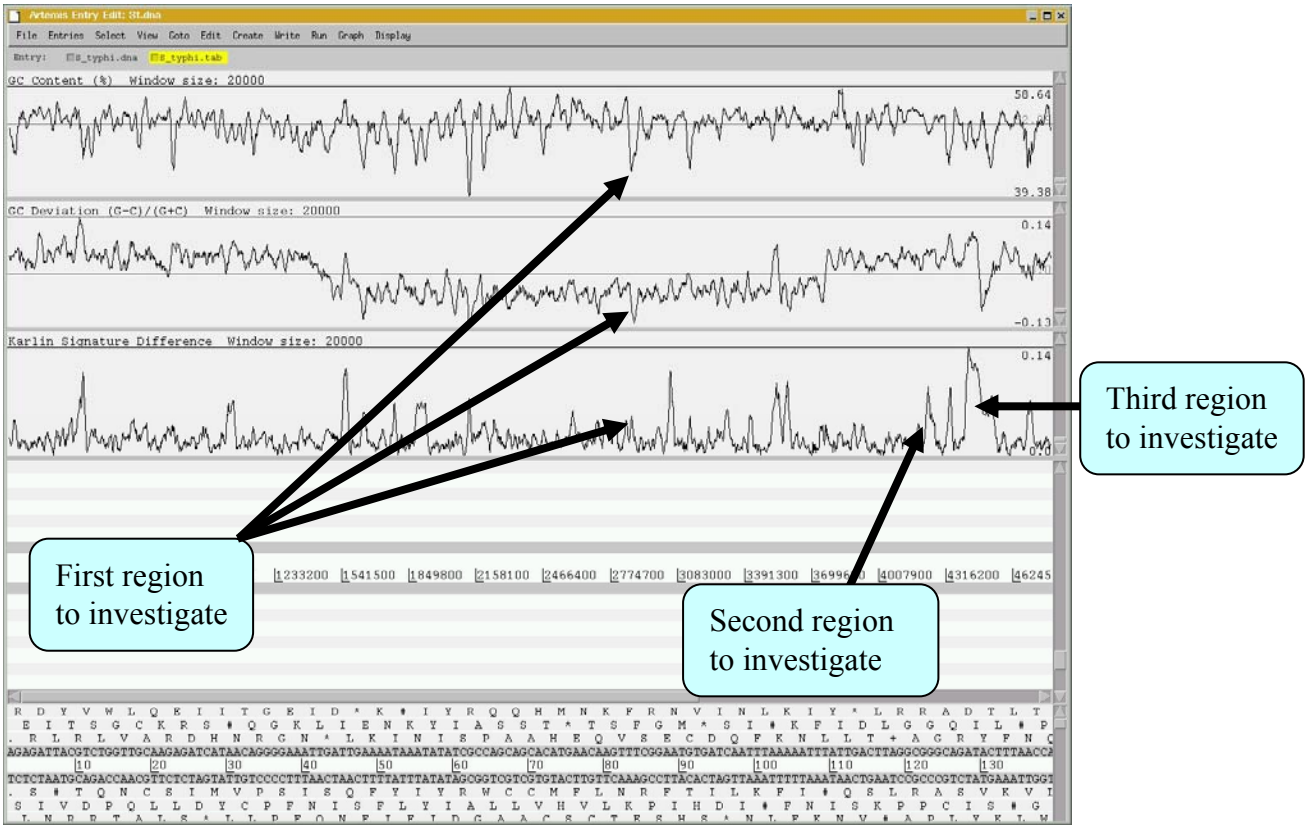
One final tip is to adjust the scaling for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window. A menu will appear with a series of values for the maximum window size (see above), select 20000. You should do this for each graph displayed.

You are now ready to zoom out by dragging or clicking the slider indicated above. Once you have zoomed out fully to see the entire genome you will need to adjust the smoothing of the graphs using the vertical graph sliders as before to have a similar view to that shown below.

Click with the left mouse button in a graph window. A line and a number will appear. The number is the relative position within the genome (bps).

Click and drag to highlight a region on the main DNA line. Notice that the boundaries of this region should now be marked in the graph windows that your previously clicked in.

Artemis Exercise 1 Part III



There are many examples where these anomalous regions of DNA within a genome have been shown to carry laterally acquired DNA. In this part of the exercise we are going to look at several of these regions in more detail. Starting with the whole genome view, note down the approximate positions and characteristics of the three regions shown above. Remember the locations of the peaks are given in the graph window if you click with the left mouse button within it.

	Genome location	**Characteristics of DNA plots**
Region 1 :	*2,860,000 bps*	*peak - karlin, troughs for G+C and CG deviation*
Region 2 :		
Region 3 :		

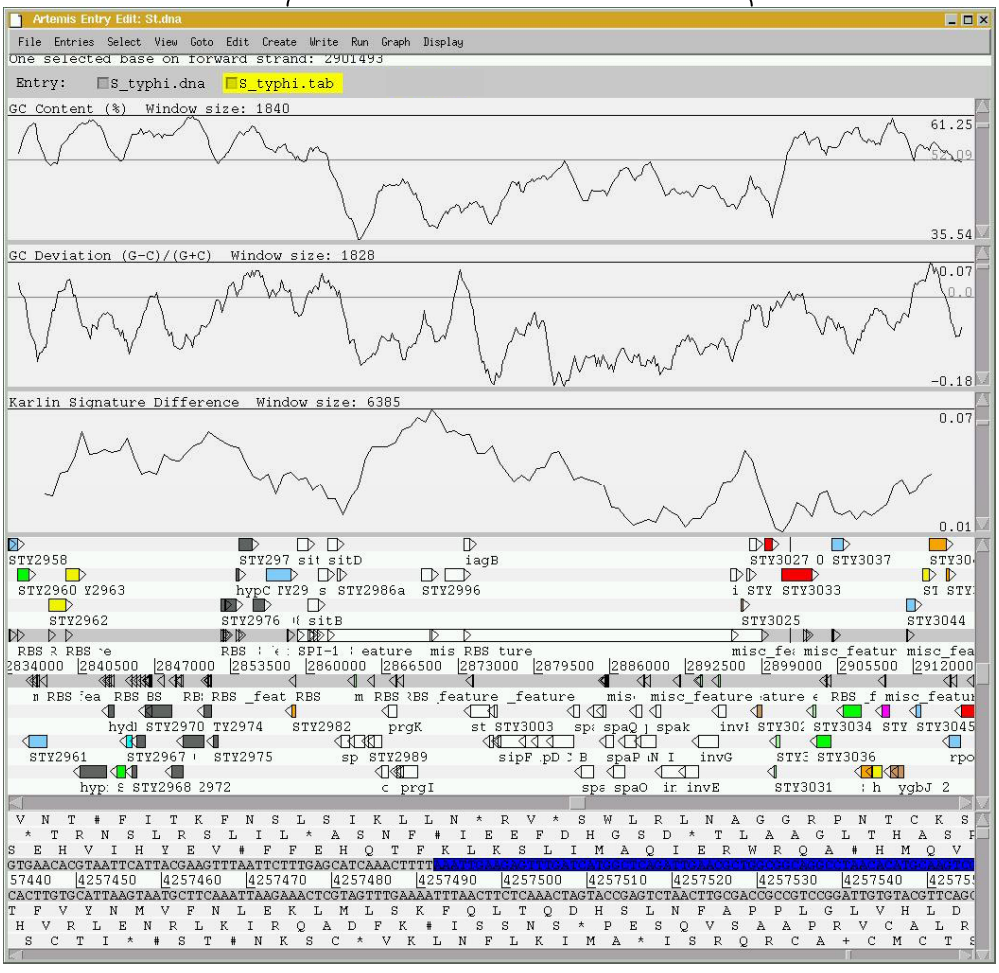
We will now zoom back into the genome to look in more detail at the first of these three peaks. Zoom into this position by first clicking on the DNA line at approximately the correct location. If you then use the vertical side slider to zoom back in, Artemis will go to the location you selected. Remember that in order to see the CDS features lying within this region you will need to turn the annotation (*S*_typhi.tab) entry back on.

The region you should be looking at is shown below and is a classical example of what is referred to as a *Salmonella* pathogenicity island (SPI). The definitions of what actually constitutes a pathogenicity island are quite diverse. However, below is a list of characteristics which are commonly seen within these regions, as described by Hacker et al., 1997.

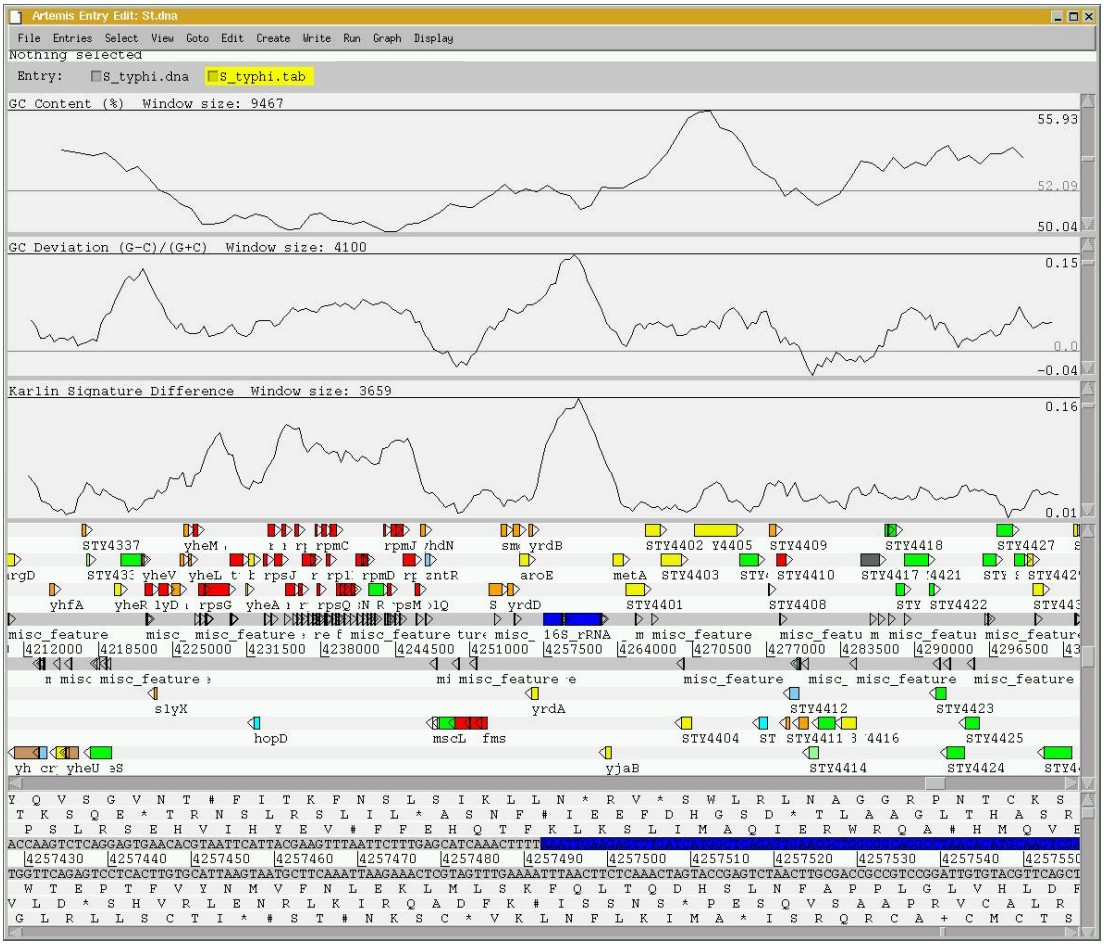
1. Often inserted alongside stable RNA's
2. Atypical G+C contents.
3. Carry virulence-related functions
4. Often carry genes encoding transposase or integrase-like proteins
5. Unstable and self-mobilisable
6. Of limited phylogenetic distribution

Have a look in and around this region and look for some of these features.

Region 1 SPI-1



Region 2



Use one of the methods you have already used to take you to the second region of interest that you noted down.

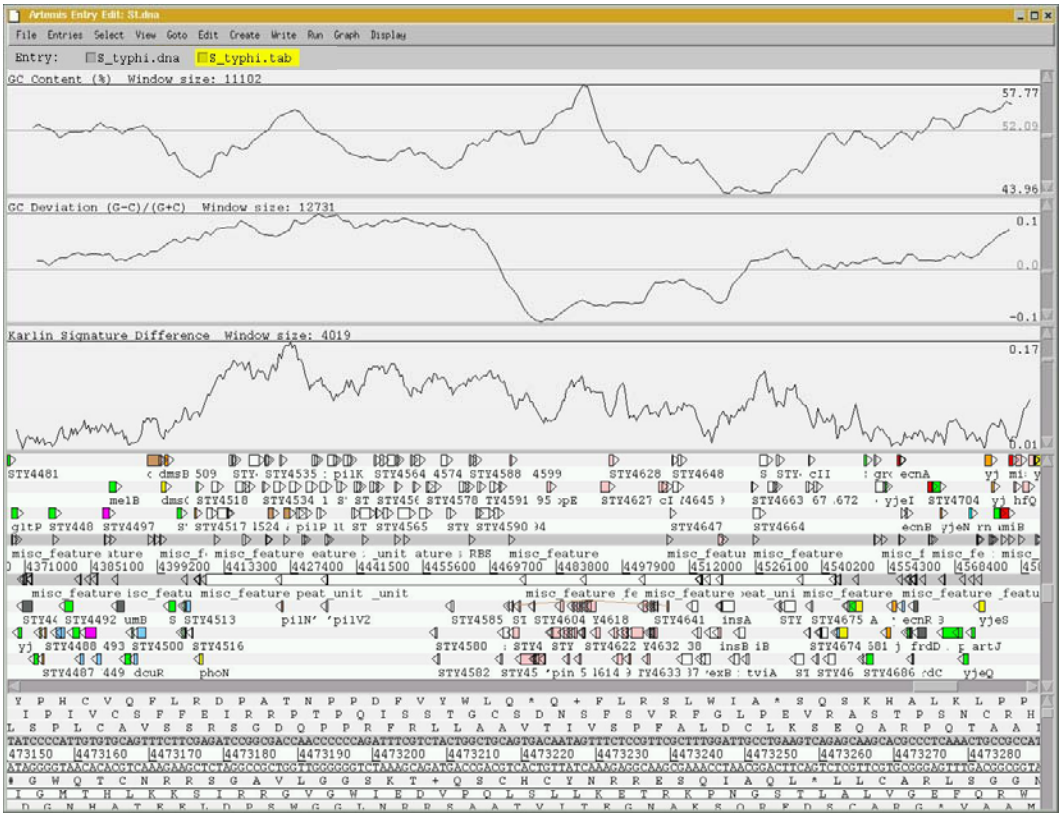
Region two acts as a cautionary note when looking at anomalous regions within a genome. Have a look at the CDSs within this region.

Does this region:

- have any of the characteristics of pathogenicity island
- are the genes within this region essential or dispensable.

Is it possible that the atypical base composition of this region is not a consequence of having originated from a foreign host. The base composition may actually be reflective of the tight sequence constraints under which this region has been maintained, in contrast to the background level sequence variation in the rest of the genome.

Region 3

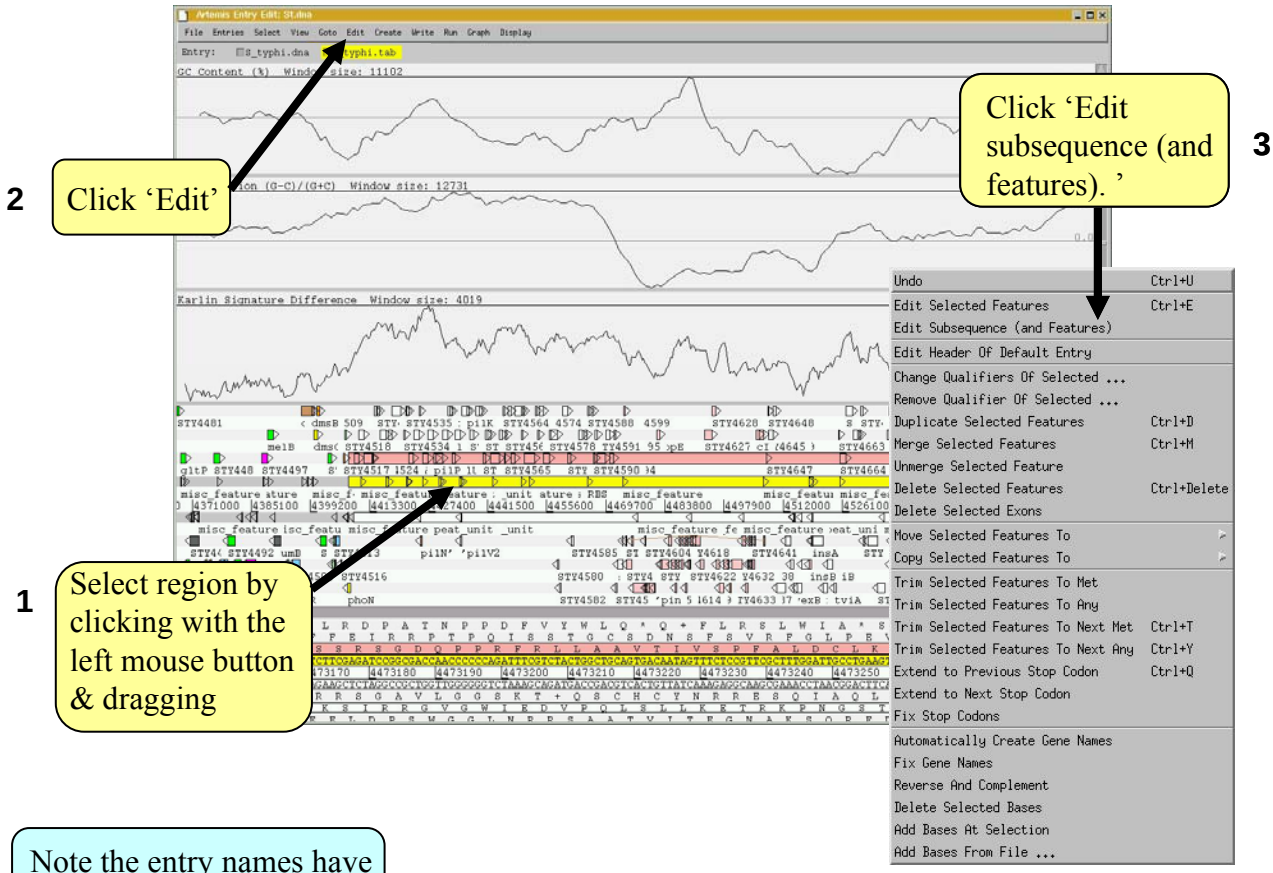


Go to region 3 as before.

Like region 1, this region is also referred to as a *Salmonella* pathogenicity island (SPI). SPI-7, or the major Vi pathogenicity island, is ~ 134 kb in length and contains ~30 kb of integrated bacteriophage. Have a look at the CDSs within this region. As before notice any stable RNAs that may have acted as the phage integration site.

Artemis Exercise 1 Part IV

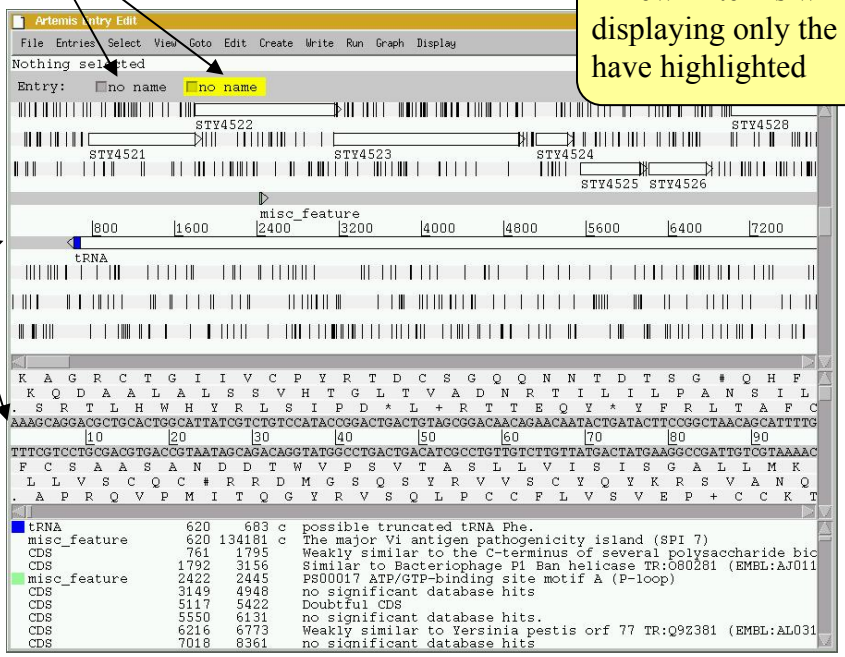
Continuing on from the analysis of Region 3 or SPI-7 (the major Vi-antigen pathogenicity island) we are going to extract this region from the whole genome sequence and perform some more detailed analysis on it. We will aim to write and save new EMBL format files which will include just the annotations and DNA for this region.



Note the entry names have changed

4

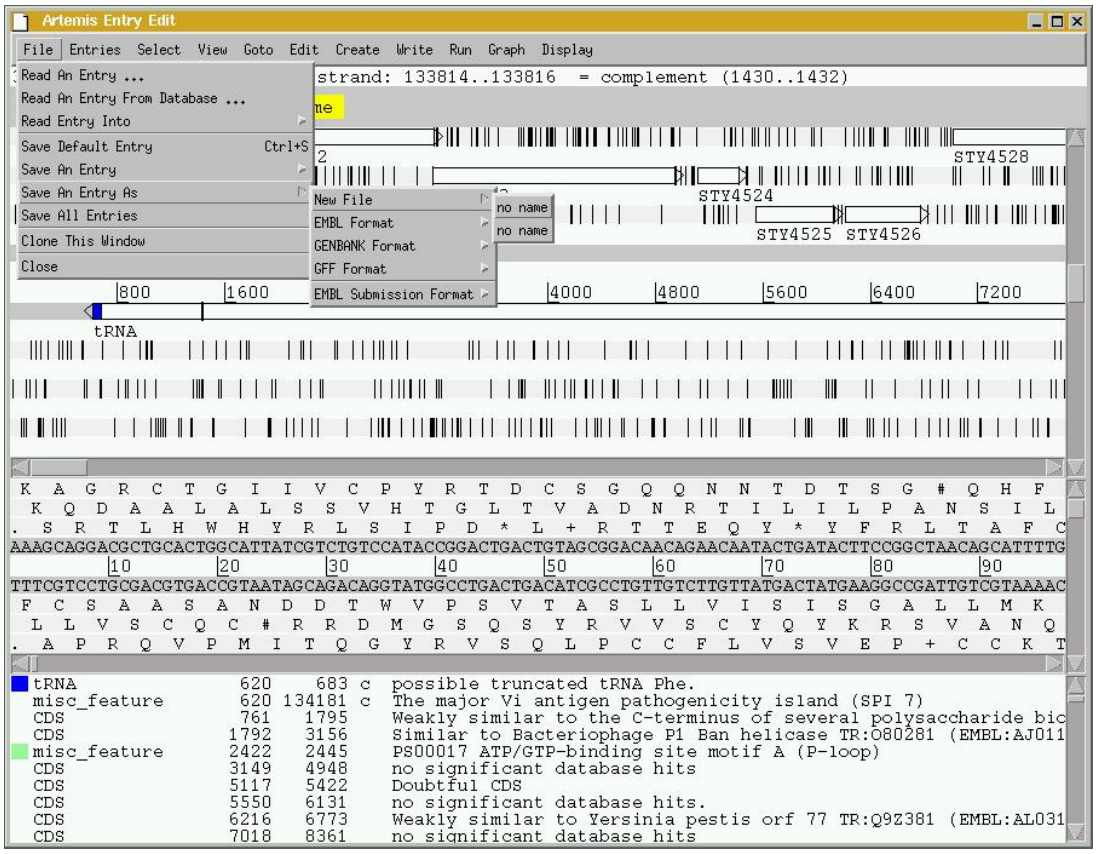
A new Artemis window will appear displaying only the region that you have highlighted



Note the bases have been renumbered from the first base you selected.

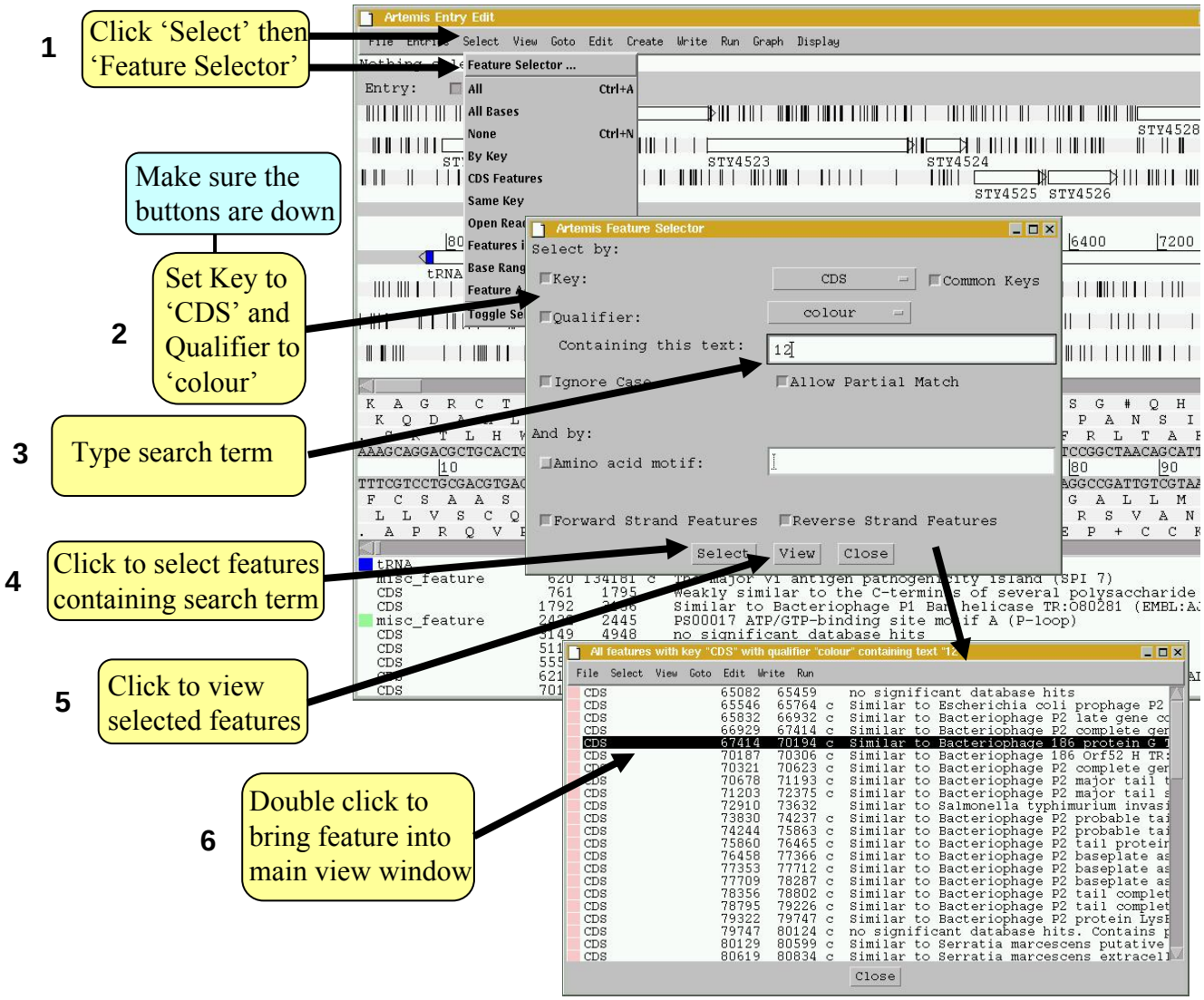
Note that the two entries on the grey Entry line are now denoted ‘no name’, they represent the same information in the same order as the original Artemis window but simply have no assigned name. Because the sub-sequence is now viewed in a new Artemis session, this prevents the original files from being over written (i.e. S_typhi.dna and S_typhi.tab). We will now save them as new files to avoid confusion. So click on the File menu then ‘Save an entry as’ and then ‘New file’. Another menu will ask you to choose one of the entries listed. At this point they will both be called ‘no name’. Left click on the top entry in the list. A window will appear asking you to give this file a name. Save this file as spi7.dna

Do the same again for the other unnamed entry and save it as spi7.tab



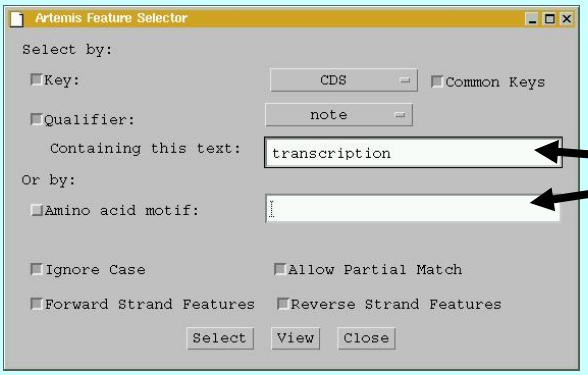
We are going to look at this region in more detail and to attempt to define the limits of the bacteriophage that lies within this region. Luckily for us all the phage-related genes within this region have been given a colour code number 12 (pink). We are going to use this information to select all the relevant phage genes using the Feature selector as shown below and then to define the limits of the bacteriophage.

First we need to create a new entry (click ‘Create’ then ‘New Entry’). Another entry will appear on the entry line called, you guessed it, ‘no name’. We will eventually copy all our phage-related genes into here.



The genes listed in **6** are only those fitting your selection criterion. They can be copied or moved in to a new entry so we can view them in isolation from the rest of the information within spi7.tab.
Firstly in window **6** select all of the CDS shown by clicking on the 'select ' menu and then selecting 'All'. All the features listed in window **6** should now be highlighted. To copy them to another entry (file) click 'Edit' then 'move selected Features To' then 'no name'. Close the two smaller feature selector windows and return to the SPI-7 Artemis window. You could rename the ' no name ' entry as you did before. Temporarily remove the features contained in 'spi7.tab' file by left clicking on the entry button on the grey entry line. Only the phage genes should remain.

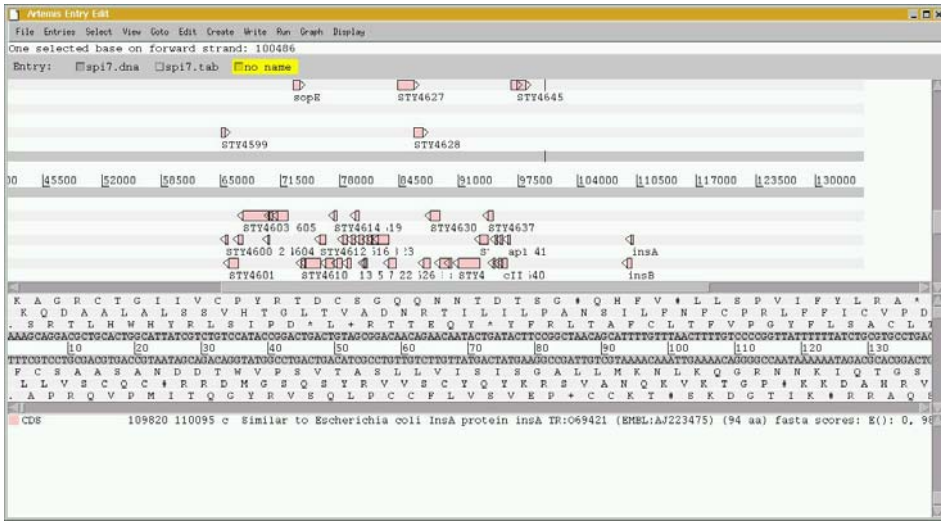
Additional methods of selecting/extracting features using the Feature Selector
It is worth noting that the feature selector can be used in many other ways to select and extract subsets of features from the genome. If you have a closer look at the Feature selector you will also see that you can use search terms to select a class or all those features with a particular amino acid motif.



Space for a search term or amino acid motif

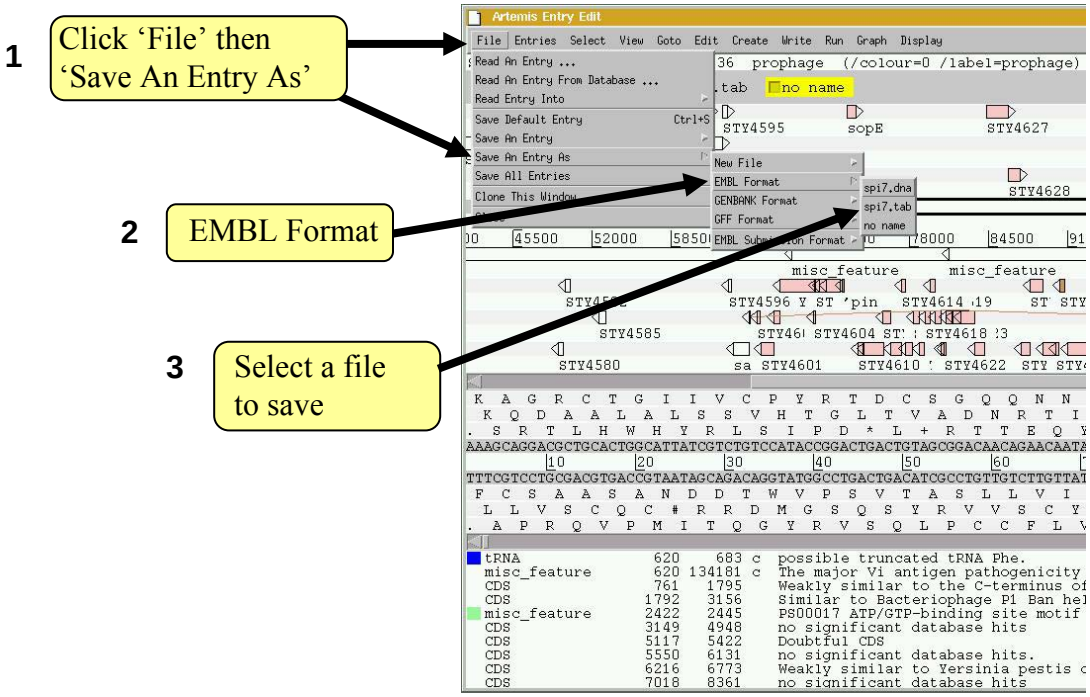
Defining the extent of the prophage.

Even from this very cursory analysis it is clear from the selection that the prophage occupies a fairly discrete region within SPI-7 (see below). It is often useful to create a DNA feature to define the limits of this type of genome landmark. To do this use the left mouse button to click and drag over the region that you think defines the prophage. Click on the create menu and select 'Create feature from base range'. A feature edit window will appear. The default ' key ' value given by Artemis when creating a new feature is 'CDS'. With this ' key ' the newly created feature would automatically be put on the translation line. However, if we change this it to 'misc_feature' (an option in the key menu top left hand corner at the edit window) Artemis will place this feature on the DNA line. This is perhaps more appropriate and is easier to visualise. If you also add in a qualifier, such as ' /label ' and add text following the /label= ????, then click ok. That text will be used as a feature label to be displayed in the main sequence view panel.



To see how well you have done turn back on the spi7.tab and have a look at the genes located at either side of your selection. Go to and look at the CDS *samA.* In reality this gene was disrupted by the insertion of this bacteriophage. If you look at the FASTA results for this CDS you may be able to track the bases between which this phage inserted.

Your final task is to write out these files in EMBL format and create a merged annotation and sequence file in EMBL format:



This will create two files one with the sequence and the other with the annotation in the directory within which you started Artemis. To create a complete EMBL file use the UNIX you covered earlier and 'cat' the files together

Artemis Exercise 2

This exercise will look at a section of the Malaria genome. You will need to close down the last Artemis exercise if you haven't already done so. Then start a new Artemis Session, as before, using the file 'Malaria.embl' in the current directory (Module_2_Artemis). Unlike the Salmonella exercise, in this instance the annotation and sequence are contained within the same file 'Malaria.embl'
The sequence you are going to look at is a small region of contrived sequence (~21 kb) taken from *Plasmodium falciparum* chromosome 13. You will see 7 CDSs, some with multiple exons. As a gentle introduction to splicing we would like you to look at the genes named , PF13_0119, MAL13P1.294 and PF13_0061. They have only been partially characterised and may in fact be missing exons. Have a look at these CDSs and confirm, edit or dismiss the proposed gene models by using G+C content, database searches and looking for splice sites (**Appendix IX**).

G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.

Note, we will cover the principals and methods of gene prediction in much more detail in a module 3.

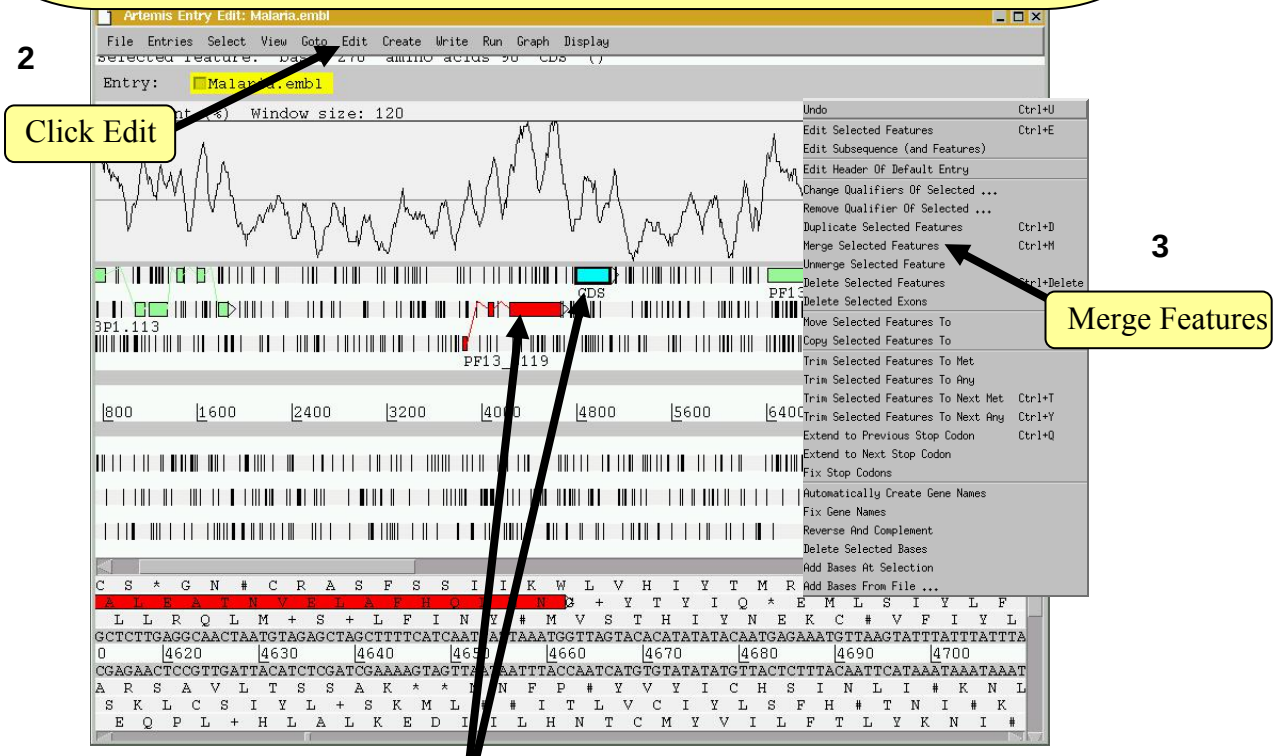
To compare the three CDS with others currently in the public databases run a fasta search. Left click the CDS, click on the 'Run' menu and then 'Run fasta on selected features'. When the search is finished, a banner will appear saying 'fasta process completed' (see above). The search may take a couple of minutes to run.

To view the search results click 'View' then 'Search Results' then 'fasta results'. The results will appear in a scrollable window. You could also view these results in your Netscape Browser window as in the previous exercise.

How does your predicted gene model for this CDS compare with proteins pulled out of the public databases? Is it possible that there are additional exons not featured in the current model.

If you think that there are additional exons that should have been included in the gene model you should add them to it. Using GC content and results from your database search as guides roughly draw in where you think the additional exon(s) lie:

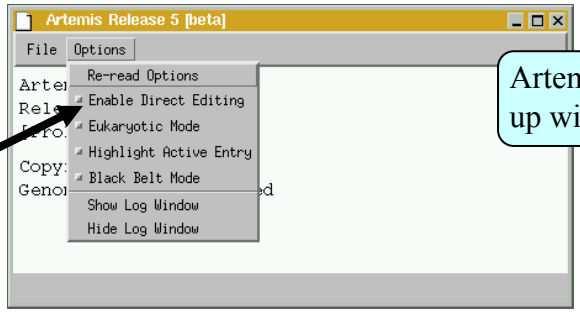
To create additional exons:
Select the region you think represents the exon by holding down the left mouse button and dragging the curser over the region of interest. Then click the 'Create' menu and select 'Create feature from base range'. A new blue CDS feature will appear on the appropriate frame line (See below).



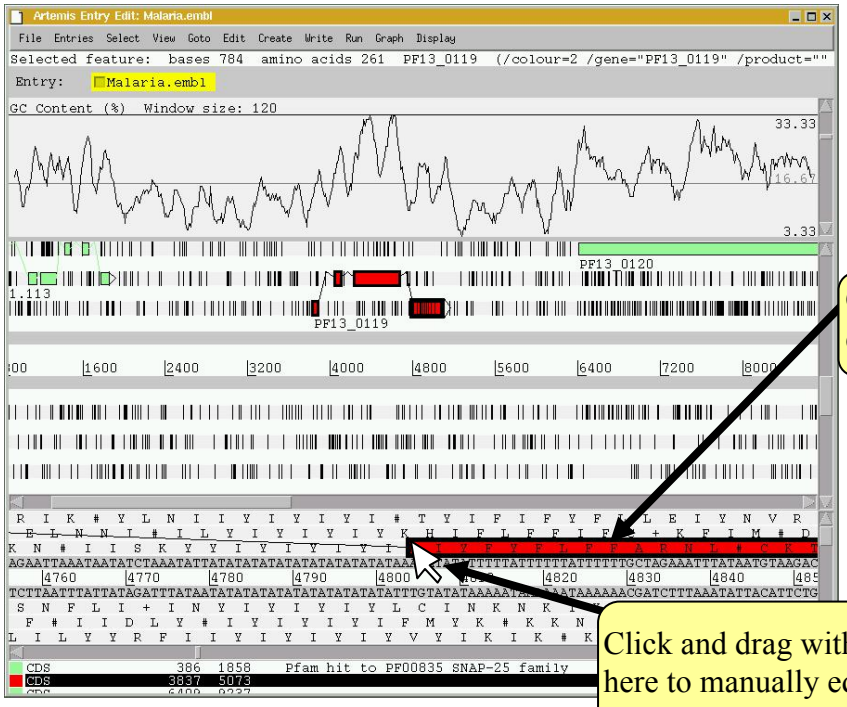
Tip, to select more than one feature (of any type) you must hold the shift key down.

The new CDS feature can then be merged with the original gene model as shown above.

A small window will appear asking you whether you are sure you want to merge these features. Another window will then ask you if you want to 'delete old features'. If you click 'yes' the CDS features you have just merged will disappear leaving the single merged CDS. If you select 'no' all of the three CDS features (the two CDSs that you started with plus the merged feature) will be retained.



You may noticed after you performed the merge function that one of the exons has subsequently jumped into another reading frame. Artemis automatically splices the CDS and so if the exon boundaries have an additional partial codon then any following exon will be pushed into another reading frame to account for this. To correct this you can edit the exon boundaries directly by turning on manual editing in the options menu of the Artemis start-up window, (as shown above). This will now allow you to edit the start and end positions of the feature boxes by using the left mouse button. Click and hold down the curser over the first or last base of any feature and then drag the mouse. The feature box should move as you drag it (see below. This can be a little tricky so please ask)



When manually editing your exons you can should look out for appropriate splice donor and acceptor sites. See below for a small list and **Appendix IX** for details of known acceptor and donor motifs for Malaria splice sites.

Once you are happy with your newly created exon re-run the fasta search and see how this compares with the other hits in the public databases. If there are more exons to mark up try and complete the gene model.

The three example CDS to analyse were selected because they have very good database hits. This obviously makes the task of making the gene model far easier. However, several of the other CDS in this region have no significant database hits. If you have time you may want to have a look at these too.

Module 2
Comparative Genomics

Introduction

The Artemis Comparison Tool (ACT), also written by Kim Rutherford, was designed to extract the additional information that can only be gained by comparing the growing number of genomes from closely related organisms.
ACT is based on Artemis, and so you will already be familiar with many of its core functions. ACT, is essentially composed of three layers or windows. The top and bottom layers are mini Artemis windows (with their inherited functionality), showing the linear representations of the genomes with their associated features. The middle window shows red blocks, which span this middle layer and link conserved regions within the two genomes, above and below. Consequently, if you were comparing two identical genome sequences you would see a solid red block extending over the length of the two sequences in this middle layer. If insertions were present in either of the genomes, they would show up as breaks between the solid red conserved regions. Data used to draw these red blocks and link conserved regions is generated by running pairwise BlastN or tBlastX comparisons of the genomes (details of how this is done are outlined in Appendix II and can be obtained from the ACT user manual: http://www.sanger.ac.uk/Software/ACT /manual/).

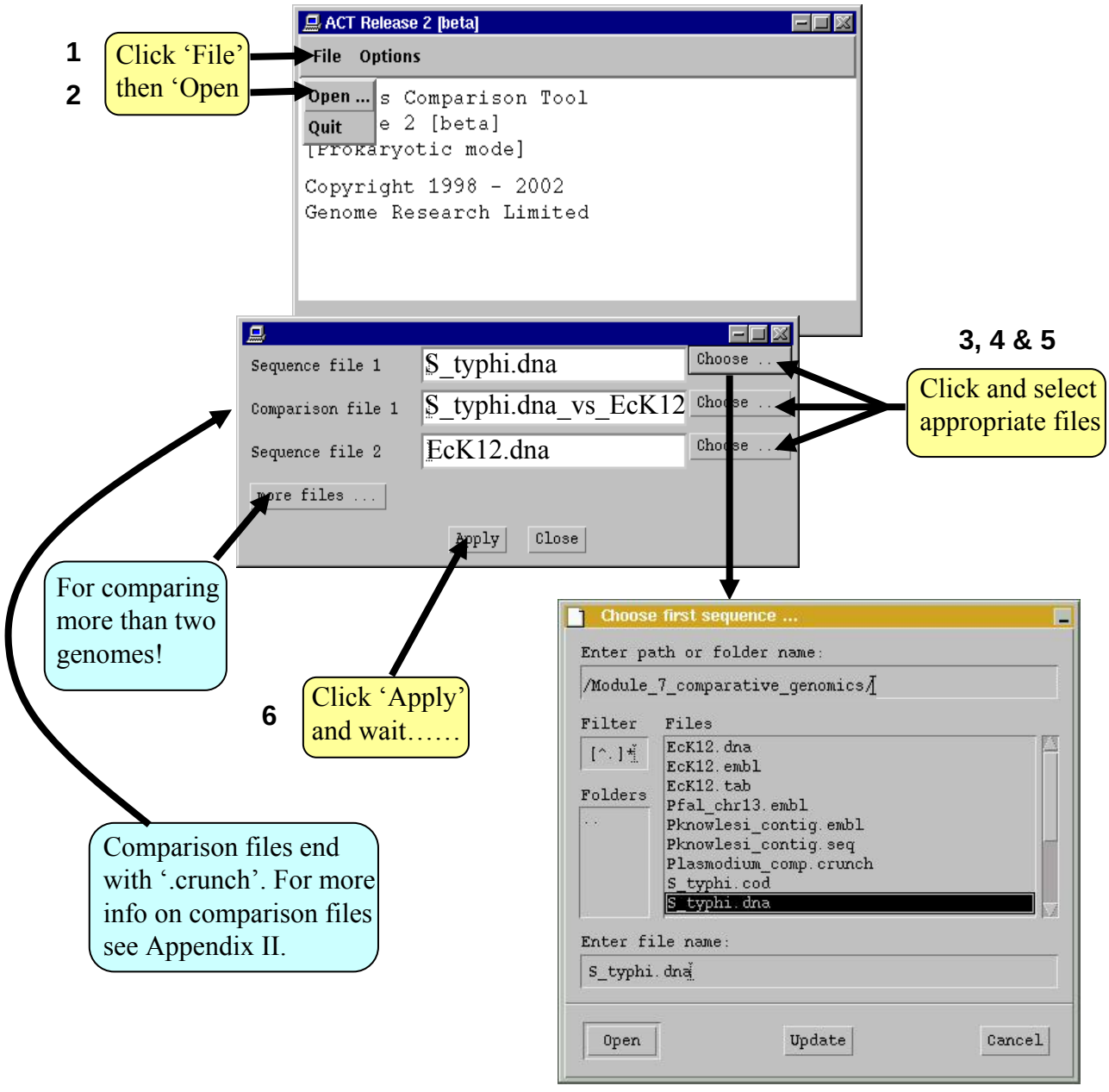
Aims

The aim of this Module is for you to become familiar with the basic functioning of ACT by using a series of worked examples. Some of these examples will touch on exercises that were used in previous Modules, this is intentional. Hopefully, as well as introducing you to the basics of ACT this Module will also show you how ACT can be used for not only looking at genome evolution but also to backup, or question, gene models and so on.

1. Starting up the ACT software

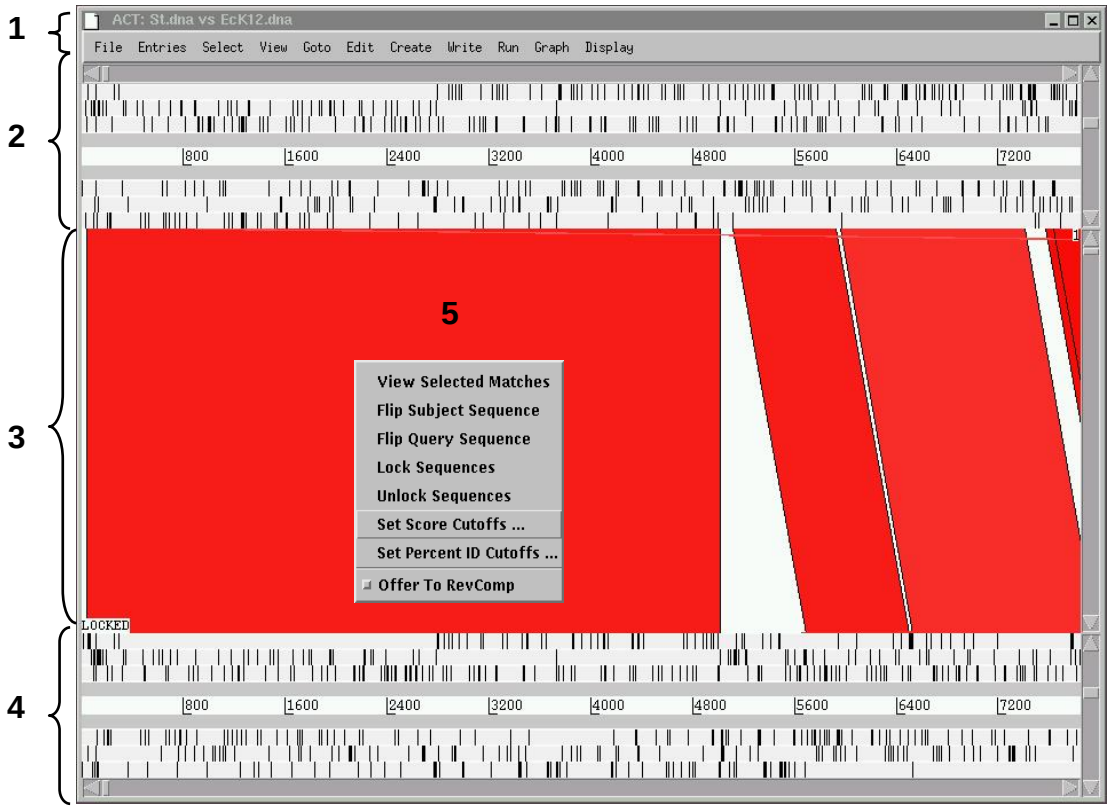
Make sure you're in the correct directory **Comparative Genomics Module 5**.
Then type
act & [return]
A small start up window will appear.
Now let's load up a *S. typhi* versus *Escherichia coli* comparison.

The files you will need for this exercise are: *S_typhi.dna*
S_typhi.dna_vs_EcK12.dna.crunch
EcK12.dna



2. The basics of ACT

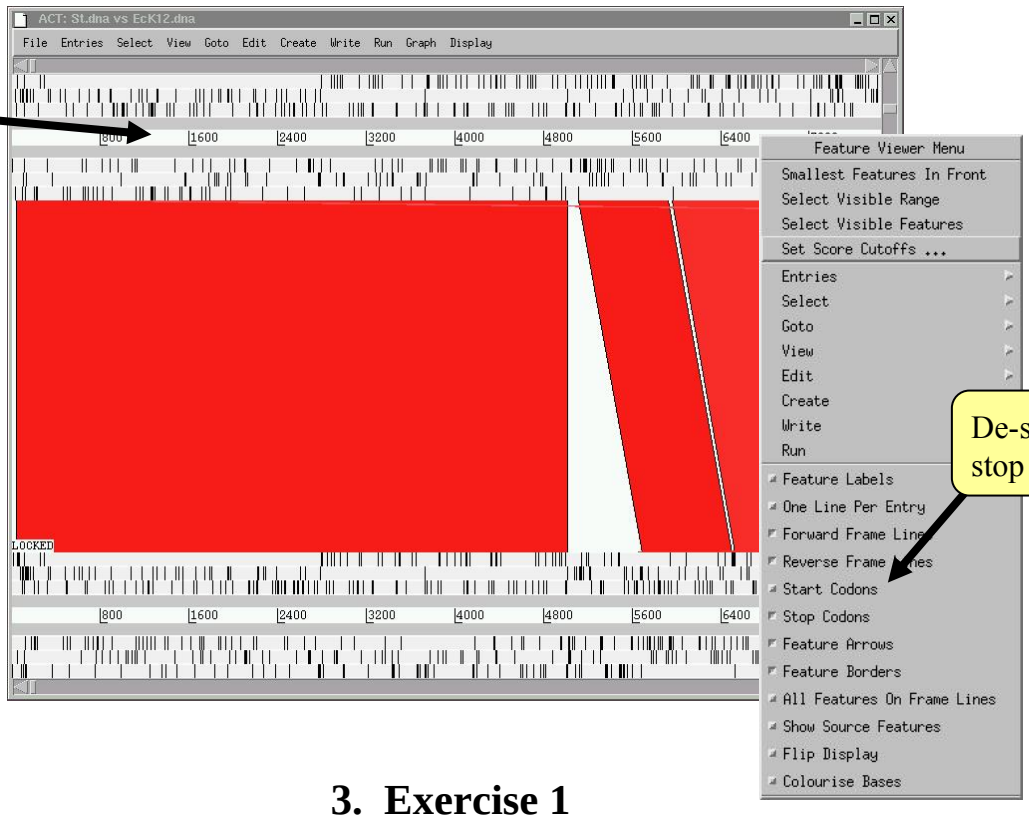
You should now have a window like this so let's see what's there.



1. Drop-down menus. These are mostly the same as in Artemis. The major difference you'll find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you selected earlier. It's a slightly compressed version of the Artemis main view panel. The panel retains the sliders for scrolling along the genome and for zooming in and out.
3. The Comparison View. This panel displays the regions of similarity between two sequences. Red blocks link similar regions of DNA with the intensity of red colour directly proportional to the level of similarity. Double clicking on a red block will centralise it.
4. Artemis-style Sequence View panel for 'Sequence file 2' (Query Sequence).
5. Right button click in the Comparison View panel brings up this important ACT-specific menu which we will use later.

1

Right button click here



2

De-select stop codons

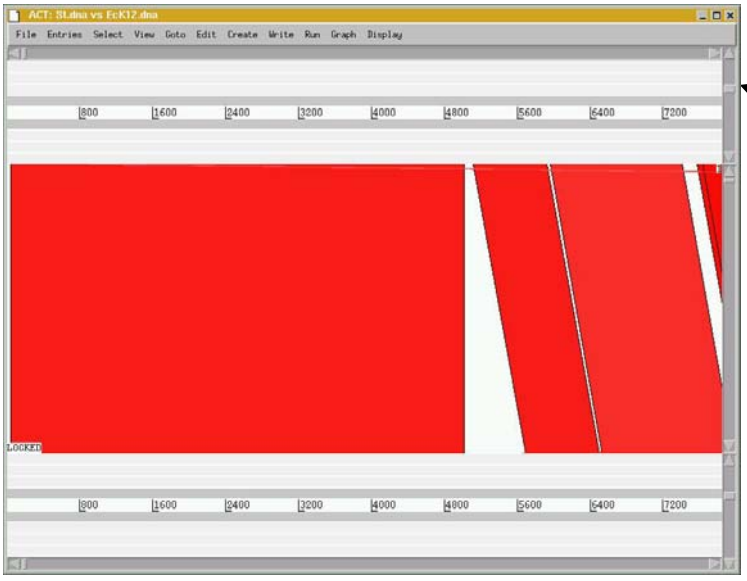
3. Exercise 1

Introduction & Aims

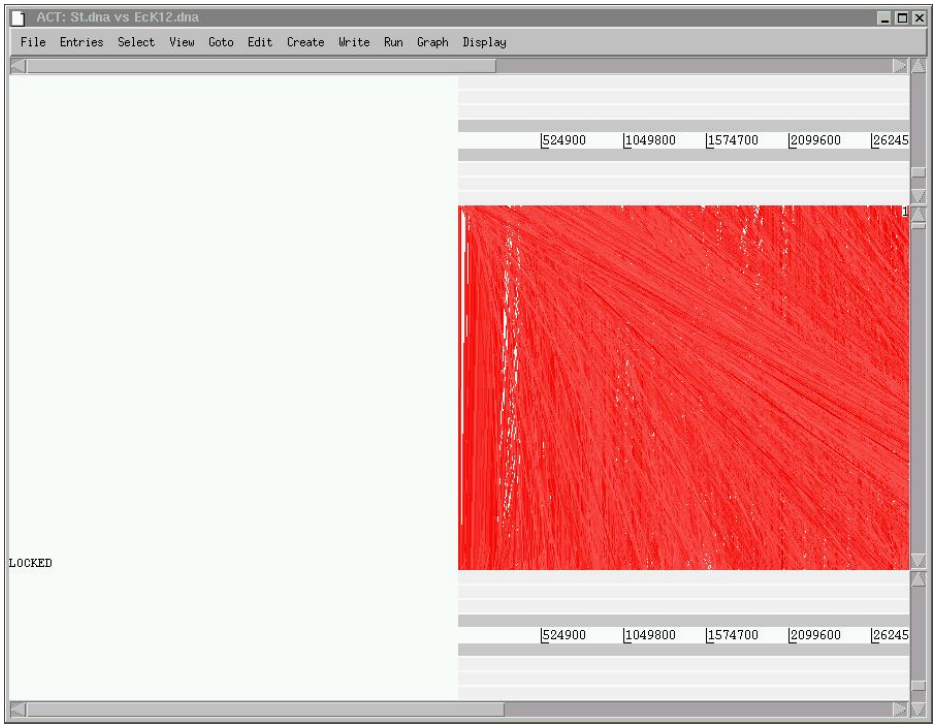
In this first exercise we are going to explore the basic features of ACT. Using the ACT session you have just opened we firstly are going to zoom outwards until we can see the entire *S. typhi* genome compared against the entire *E. coli* K12 genome. As for the Artemis exercises we should turn off the stop codons to clear the view and speed up the process of zooming out.

The only difference between ACT and Artemis when applying changes to the sequence views is that in ACT you must click the right mouse button over the specific sequence that you wish to change, as shown above.

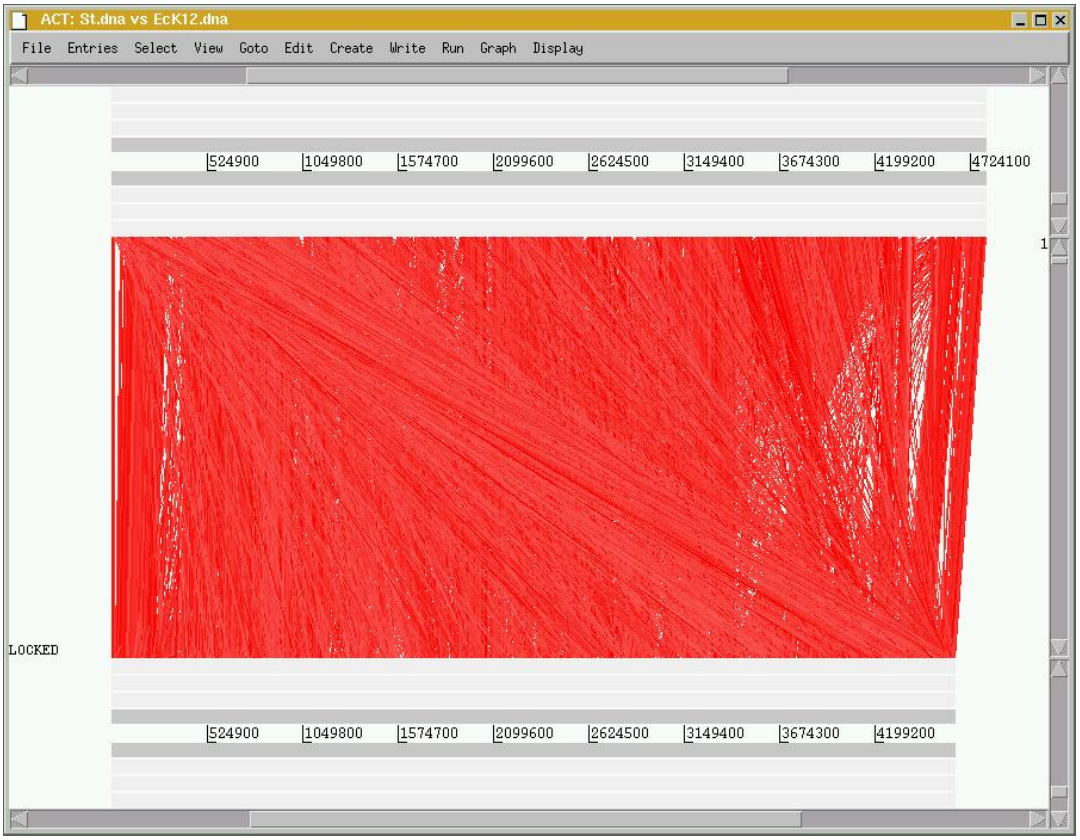
Now turn the stop codons off in the other sequence too. Your ACT window should look something like the one below:



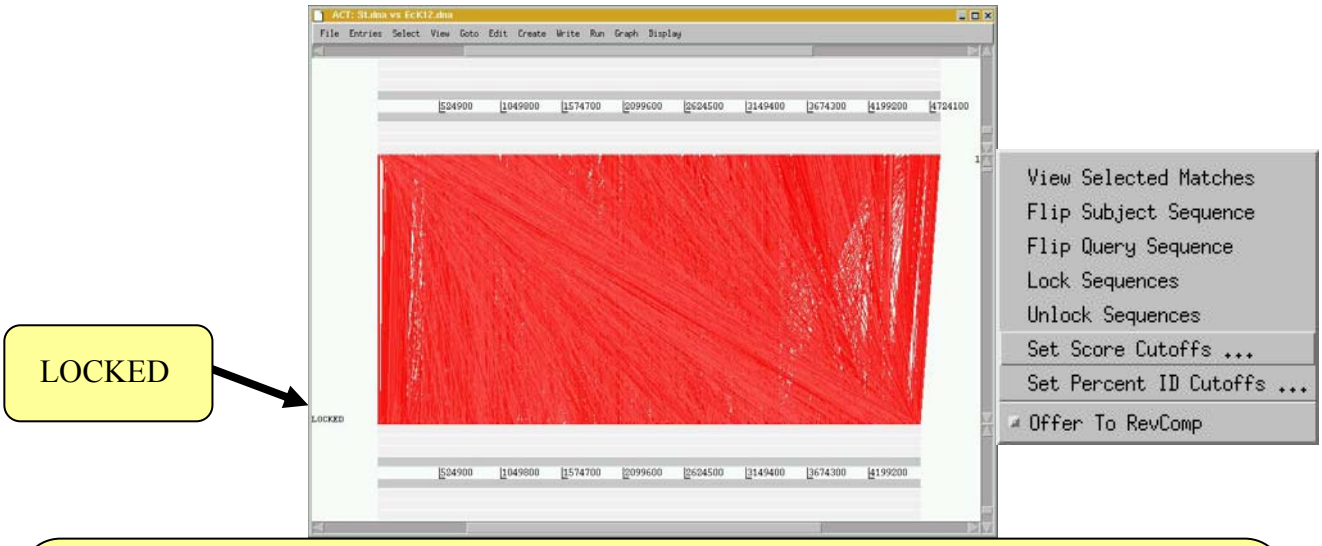
Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in synch.



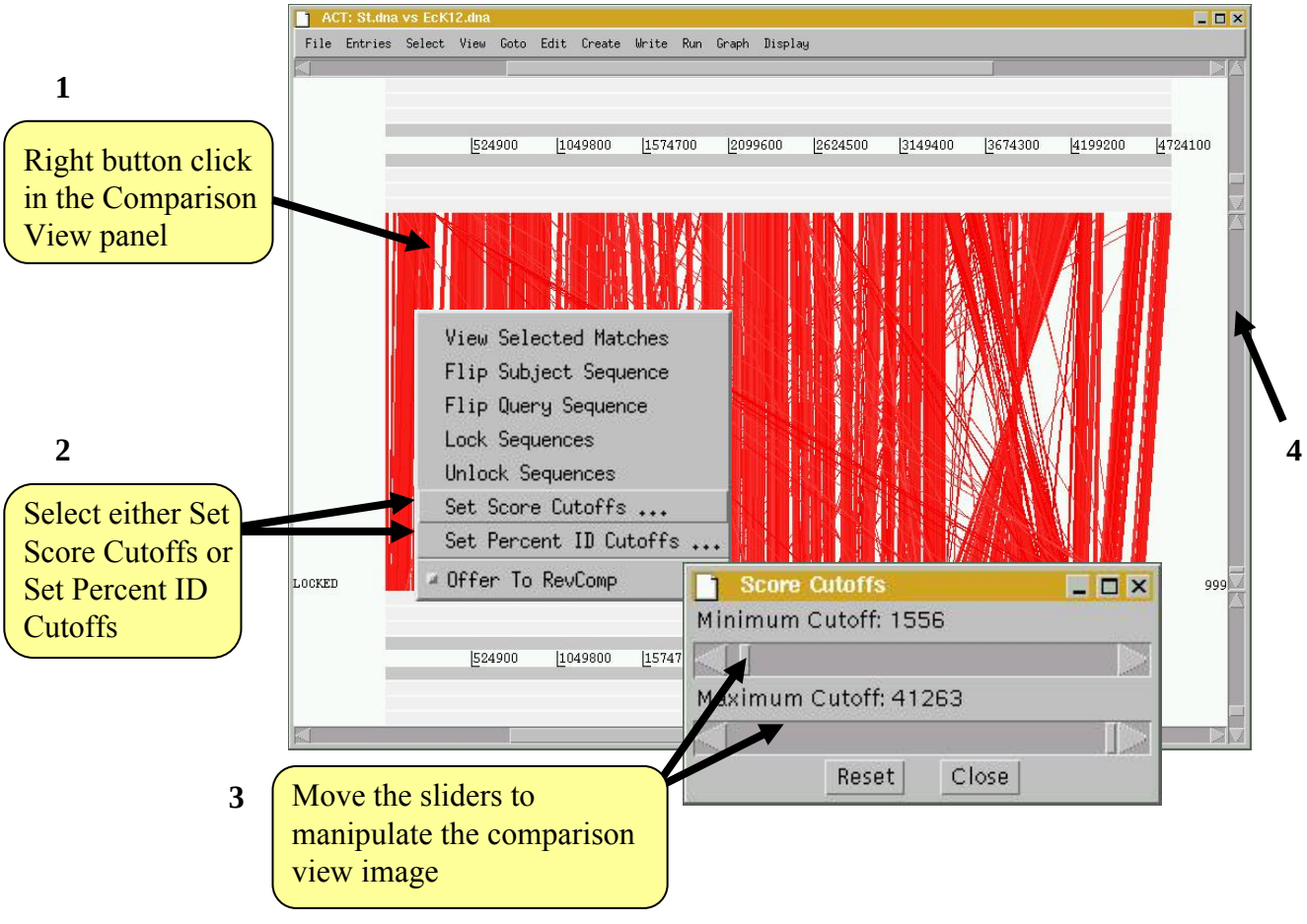
Once zoomed out your ACT window should look similar to the one shown above. If the genomes in view fall out of view to the right of the screen, use the horizontal sliders to scroll the image and bring the whole sequence into view, as shown below. You may have to play around with the level of zoom to get the whole genomes shown in the same screen as shown below.

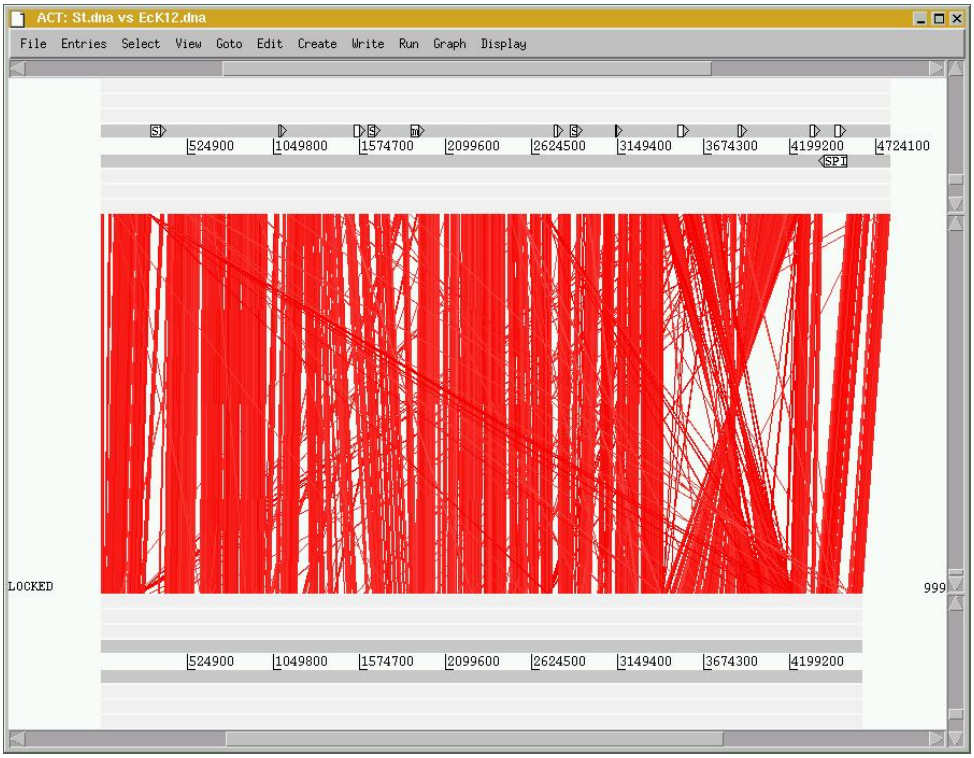


Notice that when you scroll along with either slide both genomes move together. This is because they are 'locked' together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that 'LOCKED' has disappeared from the comparison view panel and the genomes will now move independently



You can optimise your image by either removing 'low scoring' (or percentage ID) hits from view, as shown below **1-3** or by using the slider on the the comparison view panel **(4).** The slider allows you to filter the regions of similarity based on the length of sequence over which the similarity occurs, sometimes described as the "footprint".





4. Things to try out in ACT

Load into the top sequence (*S.typhi*) a '.tab' file called 'laterally.tab'. You will need to use the 'File' menu and select the correct genome sequence ('*S.typhi*.dna') before you can read in an entry. If you are zoomed out and looking at the whole of both genomes you should see the above. The small white boxes are the regions of atypical DNA covering regions that we looked at in the first Artemis exercise. It is apparent that there is a backbone sequence shared with *E. coli* K12. Into this various chunks of DNA, specific the *S. typhi* (with respect to *E. coli* K12) have been inserted.

5. More things to try out in ACT

1. Double click red boxes to centralise them.
2. Zoom right in to view the base pairs and amino acids of each sequence.
3. Load annotation files into the sequence view panels.
4. You could load in the appropriate '.tab' files for each genome (S_typhi.tab and EcK12.tab) and view the annotation of a particular region. Also try using some of the other Artemis features eg. graphs etc.
5. Find an inversion in one genome relative to the other then flip one of the sequences.

Once you have finished this exercise remember to close this ACT session down completely before starting the next exercise

6. Exercise 2 Part I:

Plasmoidum falciparum and Plasmodium knowlesi: Genome Comparison

Introduction

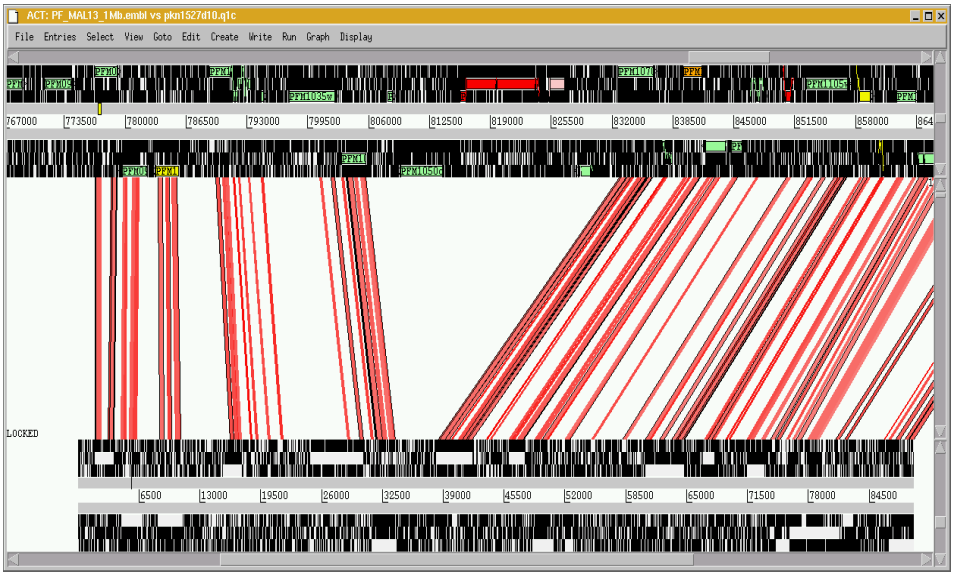
The parasite *P. falciparum* is responsible for hundreds of millions of cases of malaria and causes over 1 million deaths every year. Treatment and control have become difficult with the spread of drug-resistant malaria strains across the endemic countries in the world and there has been a major emphasis on research as part of our search for new drugs / vaccine candidates to fight against malaria. The analysis of the whole genome of *P. falciparum* has been completed and is made publicly available by the Malaria Genome Sequencing Consortium . Several animal models of malaria have also been used by researchers to study several aspects of malaria biology / host-parasite interactions. Sequences representing partial genomes of some of these model malaria parasites are also available now. This allows us to perform comparative analysis of the genomes of malaria parasites and understand the basic biology of their parasitism, based on the similarities / dissimilarities between the parasites at DNA / predicted protein level.

Aim

You will be looking at the comparison between a genomic DNA fragment of the primate malaria *P. knowlesi* and the previously annotated chromosome 13 of *P. falciparum* . By comparing the two genomic fragments you will be able to study the degree of conservation of gene order and identify new genes in *P. knowlesi* genome. As part of the exercise you will also identify any gross dissimilarity visible between the the two genomic fragments and finally, predict/ modify the gene model for one multi-exon gene in *P. knowlesi* genomic fragment.

The files that you are going to need are:

Pfal_chr13.embl	- annotation file with sequence
Pknowlesi_contig.seq	- sequence file (without annotation)
Pknowlesi_contig.embl	- annotation file with sequence
Plasmodium_comp.crunch	- tblastx comparison file



P. falciparum chr 13 (fragment)

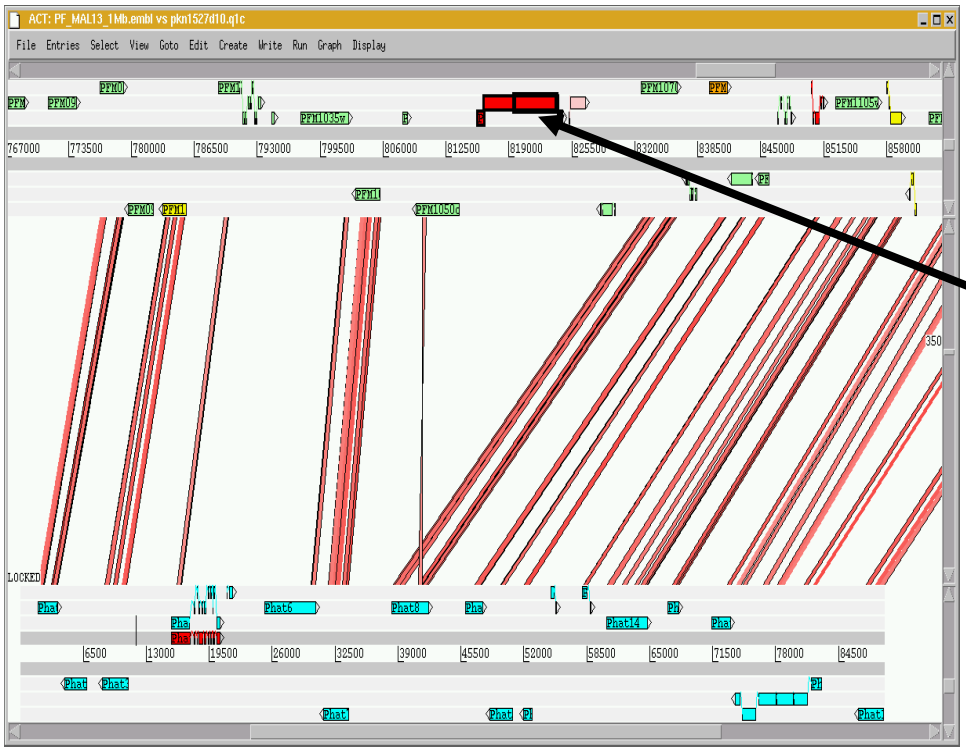
P. knowlesi contig

Comparison of *P. knowlesi* contig and the annotated chromosome 13 fragment of *P. falciparum*

Exercise 2 Part II

Conservation of gene order (synteny)

- In the ACT start up window load up the files Pfal_chr13.embl, Pknowlesi_contig.seq and the comparison file Plasmodium_comp.crunch
- Use the slider on either sequence view panel to obtain a global view of the genome comparison. Also used the slider on the comparison view panel to remove the 'shorter' similarity hits. What effects does this have?
- Can you see conserved gene order between the 2 species?
- Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within this unique region in file: Pfal_chr13.embl (top sequence)
- Example location: **Pfal_chr13.embl**, 815823..829969
- What are the predicted products of the genes assigned to this unique location? View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.
- Can you identify a few putative genes in *P. knowlesi* contig, based on their conserved and syntenic nature with *P. falciparum* chromosome 13? Activate / inactivate stop / start codons in an entry, using the right click button on the mouse. This will allow you to see any potential ORFS.
- Any thoughts about the possible biological relevance of the comparison?



P. falciparum
Pfal_chr13.embl

What is the gene product?

P. knowlesi
Pknowlesi_contig.embl

Exercise 2 Part III

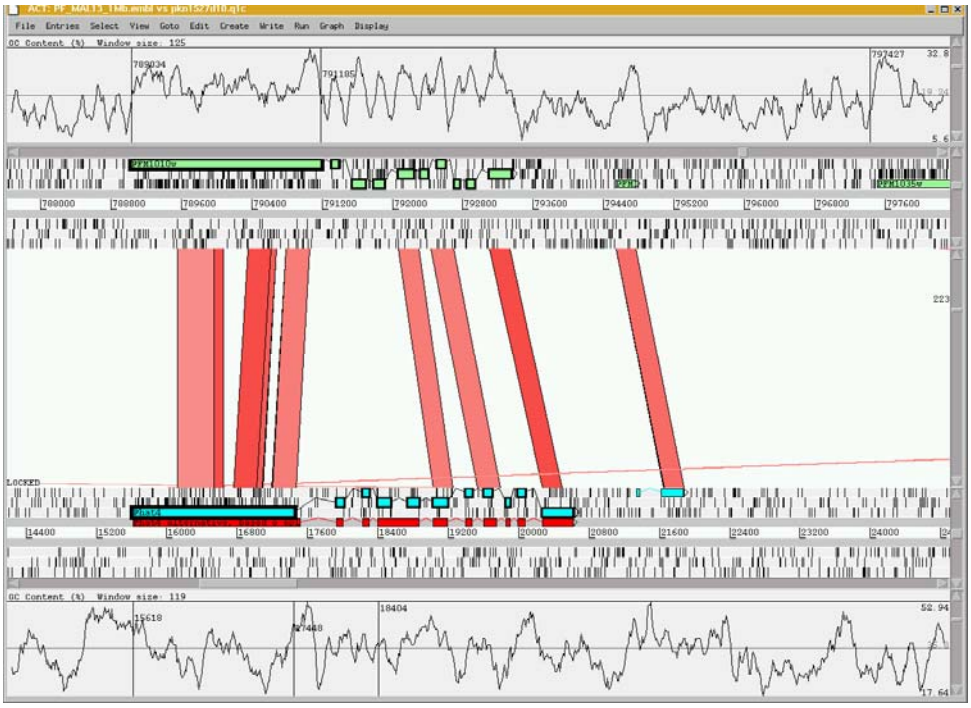
Prediction of gene models:

There are several computer algorithms covered earlier in Module 3 that predict gene models, based on training the algorithm with previously known gene sets with previously known experimentally verified exon-intron structures (in eukaryotes). However, no single programme can predict the gene structure with 100% accuracy and one needs to curate / refine the gene models, generated by automated predictions. We have generated automated gene models for the *P. knowlesi* contig, using PHAT (Pretty Handy Annotation Tool, a gene finding algorithm, see in Mol. Biochem. Parasitol. 2001 Dec;118(2):167-74) and the automated annotation is saved in Pknowlesi_contig.embl.

- Zoom into the *P. falciparum* gene labelled PFM1010w shown below. Can you compare the 2 gene models and identify the conserved exon(s) between the 2 species?
- Use the slider on the comparison view panel to include some 'shorter' similarity hits. Can you now identify all the conserved exons of the PFM1010w orthologue in the *P. knowlesi* contig? (For the time being, disregard the misc_feature for 'Phat4', coloured in red in the 'Pknowlesi_contig.embl' file)
- Open the '*GC Content (%)*' window from '*graph*' menu for both the entries. Can you relate the exon-intron boundaries to GC-content for the *P. falciparum* gene labelled PFM1010w? Is it also applicable to the gene model 'Phat4' in the *P. knowlesi* contig?
- Example regions:

  Pfal_chr13.embl, 789034..793351

  Pknowlesi_contig.embl, 15618..20618



P. falciparum
Pfal_chr13.embl

P. knowlesi
Pknowlesi_contig.embl

Comparison between orthologous genes in *P. falciparum* and *P. knowlesi*

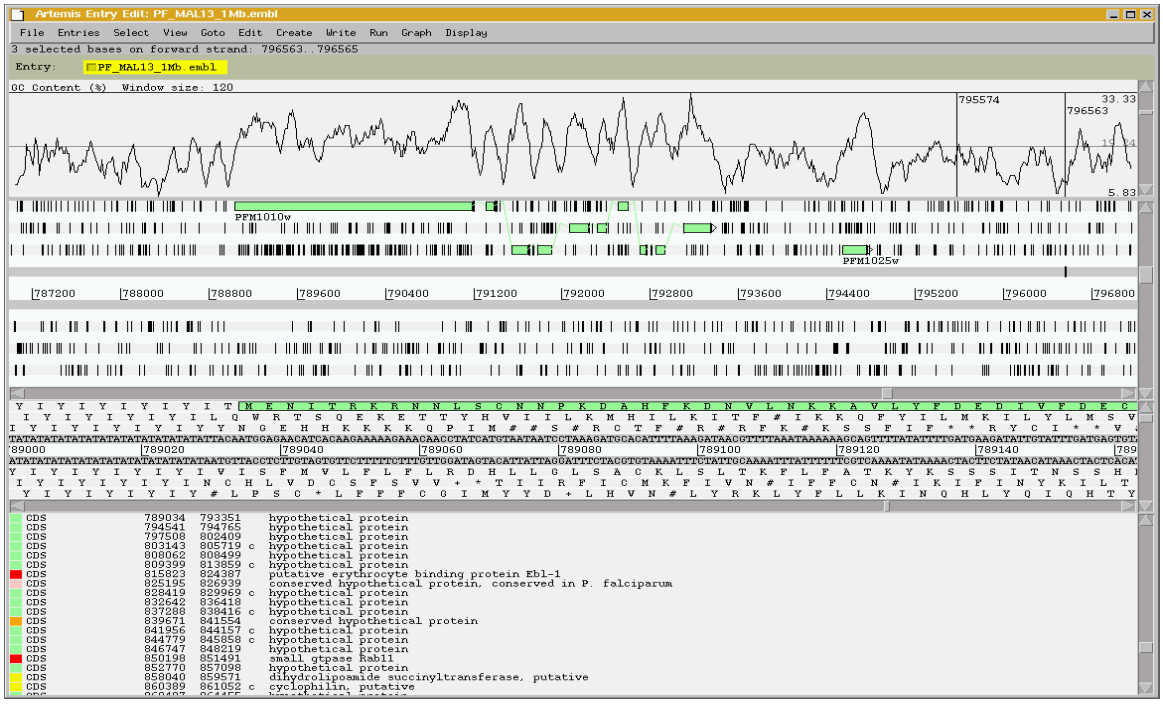
Exercise 2 Part IV

Gene models for multi-exon genes in *P. falciparum*:

- Use '*File*' menu to select entry 'Pfal_chr13.embl' and select '*Edit In Artemis'* to bring up an Artemis window.
- In Artemis window, use '*Graph*' menu and switch 'on' the '*GC Content (%)'* window.
- Use '*Goto*' menu to select '*Navigator*' window and within the Navigator window, select '*Goto Feature With This qualifier value'* and type 'PFM1010w', click then close the dialogue box.
- Go through the annotated gene model for 'PFM1010w' and have a look at the the exon-intron boundaries and compare with the splice site sequences from *P. falciparum* given in Appendix IX.
- Also have a glance through a few other gene models for multi-exon genes and have a look at the intron sequences as well. Can you find any common pattern in the putative intron sequences? Hint – look at the complexity of the sequence
- You can delete exon(s) of any gene by selecting the exon(s) and then choosing '*Delete Selected Exons'* from '*Edit*' menu. Similarly, you can add an exon to a particular gene by co-selecting the exon and the gene (CDS features) followed by selecting '*Merge Selected Features'* from the '*Edit*' menu.

- Example regions:

  Pfal_chr13.embl, 789034..793351, 657638..660023, 672361..673753



Example location: 789034..793351, in Pfal_chr13.embl

Exercise 2 Part V

Curation of gene models in *P. knowlesi*:

We are now going to edit the gene model for *P. knowlesi*.

•Use '*File*' menu from the ACT displaying *P. falciparum* and *P. knowlesi to* select entry 'Pknowlesi_contig.embl' and select '*Edit In Artemis'* to bring up an Artemis window.

•Within the Artemis window, use '*Graph*' menu and switch 'on' the '*GC Content (%)*' window.

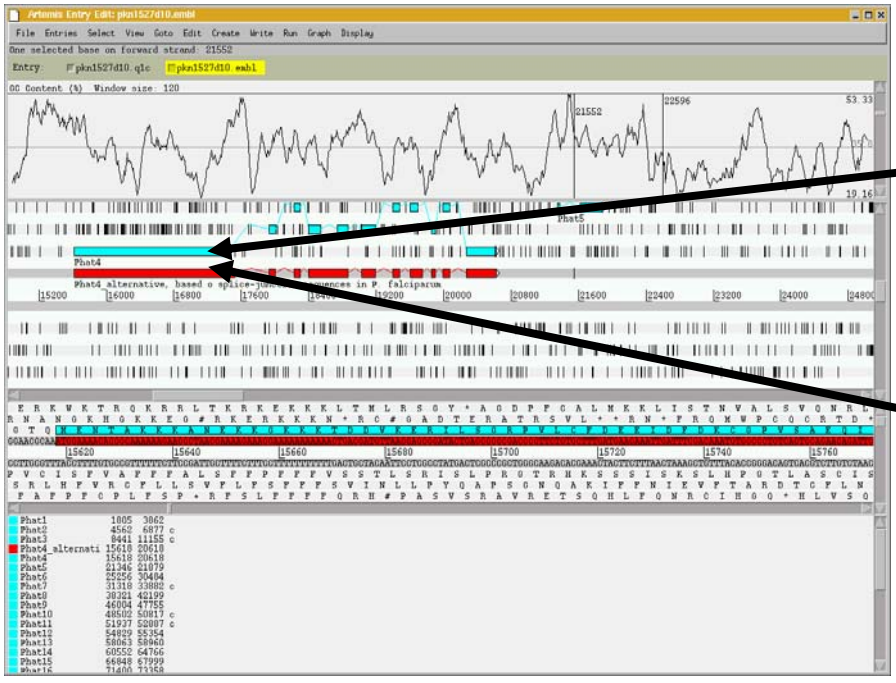
•Use '*Goto*' menu to select '*Navigator*' window and within the Navigator window, select '*Goto Feature With This Text'* and type 'Phat4'.

•Go to the first ACT window, and use the '*Options*' menu to select '*Enable Direct Editing'*

•Go through the gene model of 'Phat4' and have a glance through the exon-intron boundaries. Can you suggest any alternative gene model, after consulting the Table provided in Appendix IX, containing several examples of experimentally verified splice site sequences for *P. falciparum*?

•Example modifications:

Have a look at the '*misc_feature'*, coloured in red (location: 15618..20618). Can you spot any difference in the red gene model of 'Phat4' at the exon-intron boundaries? Select the red feature, click on '*Edit*' menu and select '*Edit Selected Features'* and in the new window that pops out, change the 'Key' from misc-feature to 'CDS' and click on 'OK' button to close the window. Now you can compare the automatically created blue gene model and the curated red gene models at protein level and predict any alternative splicing pattern.



Example location: 15618..20618, in Pknowlesi_contig.embl

Exercise 3

Introduction

Having familiarised yourselves with the basics of ACT, we are now going to use it to look at a region of synteny between *T. brucei* and *Leishmania*.

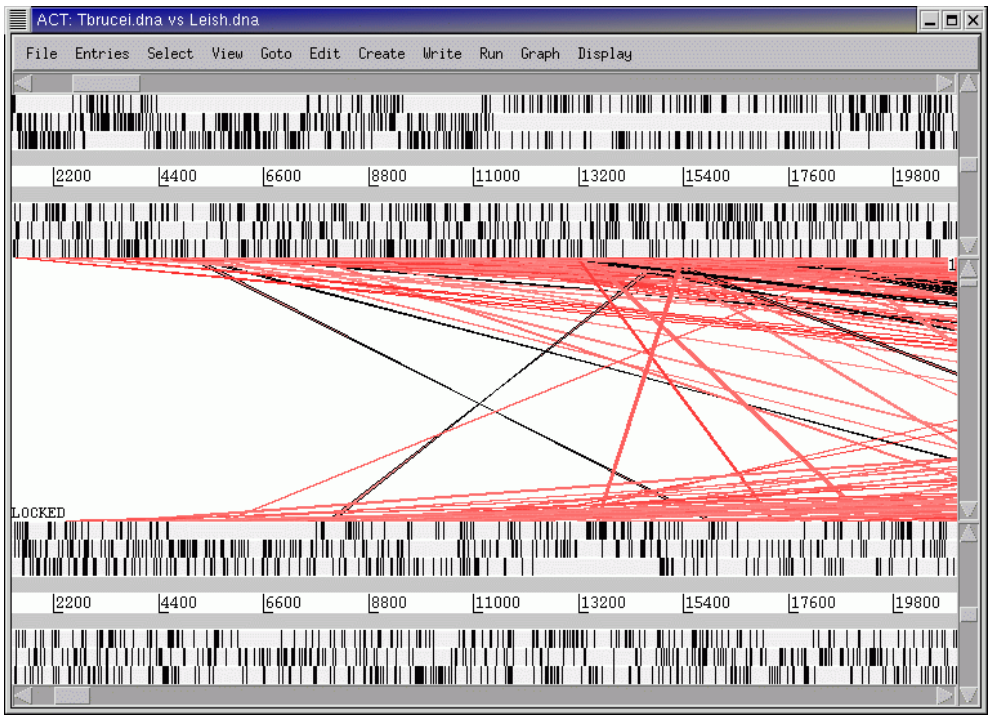
Aim

By looking at a comparison of the annotated sequences of *T. brucei* and *L.major* you will be able to analyse, in detail, those genes that are found in both organisms as well as spot the differences. You will also see how act can be used to study the different chromosome architecture of these two parasite species.

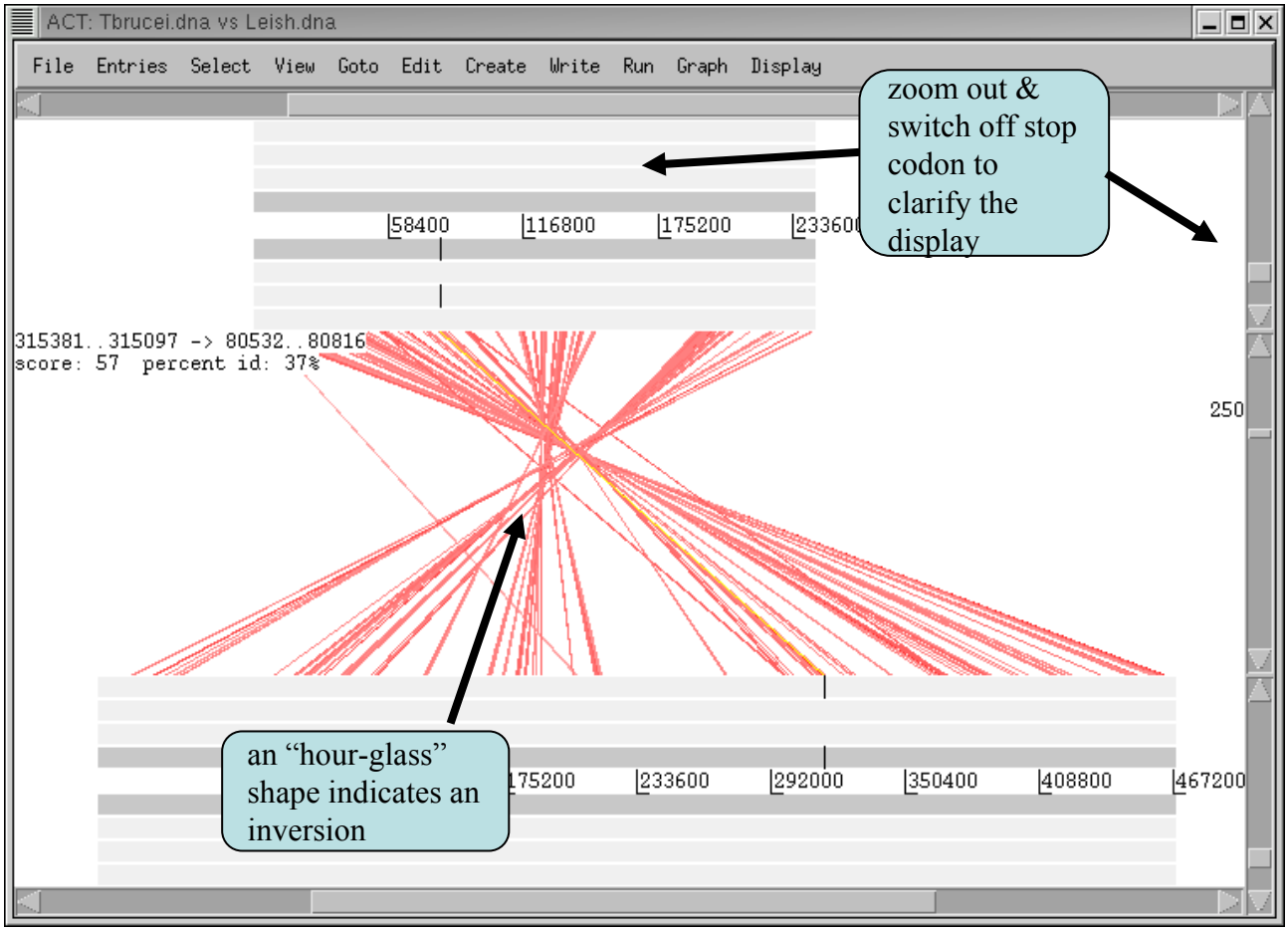
The files that you are going to need are:

Tbrucei.dna	- *T. brucei* sequence
Tbrucei.embl	- *T. brucei* annotation
Leish_vs_Tbrucei.tblastx	- comparison file
Leish.dna	- *L. major* sequence
Leish.embl	- *L. major* annotation

First, load up the sequence files for *T. brucei* and *L. major* and the comparison file in ACT.



Next, you need to find the regions of synteny between the sequences.



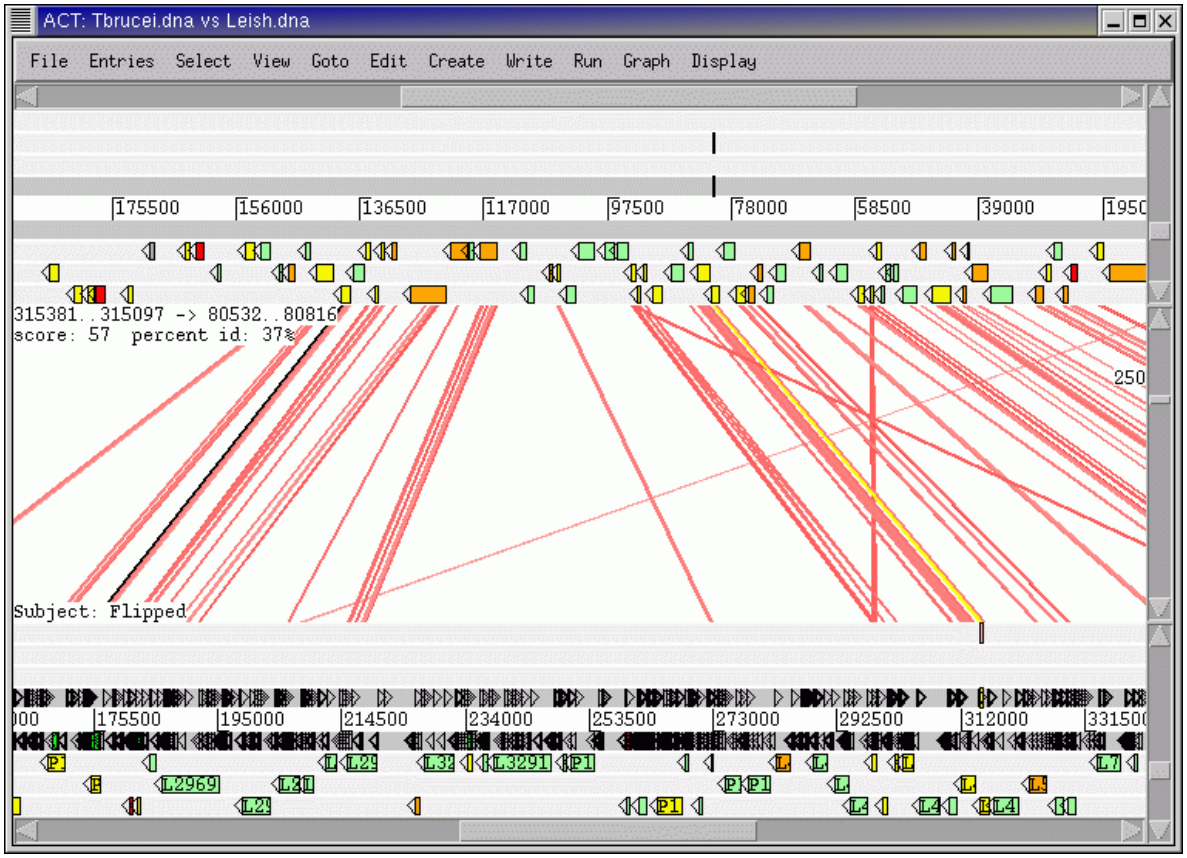
When you have determined where there is synteny, zoom in to the region for a detailed look. At this point you can add the annotation from the files called **Leish.embl** and **Tbrucei.embl**.

Can you see conserved gene order between the 2 species?

Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within theses regions.

What are the predicted products of the genes assigned to these locations? View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.

Can you identify any genes in one organism that don't appear to be predicted in the other? If so, add these to your annotation.



Exercise 4

Introduction

The quinic acid gene cluster (the *qut* cluster) is present among many filamentous fungi including including *Aspergillus fumigatus, Neurospora crassa*, *Aspergillus nudulans* and *Podospora anserina*. Although these fungi belong to the same fungal taxonomic family (Ascomycetes), they vary greatly in their biological characteristics. In this exercise you will be studying and comparing the organisation of *qut* gene cluster among these 4 fungi, using ACT.

Aim

By looking at a comparison of the annotated sequences of *N. crassa*, *A. fumigatus* and *A. nidulans* you will be able to first, add annotations to qut cluster genes in P. anserina sequence and second compare those genes that are found in all 4 organisms as well as spot the differences and study the synteny.

The files that you are going to need are:

1) N_crassa_qut.embl - sequence & annotated file for *N. crassa*
2) A_fum_qut.embl - sequence & annotation file for *A. fumigatus*
3) A_nid_qut.embl - sequence & annotation file for *A. nidulans* (artificially joined contig)
4) P_anserina_qut.embl - sequence & gene model file for *P. anserina* (without annotation)

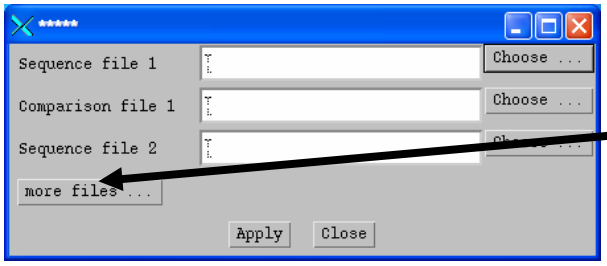
5) A_fum_N._crassa.comp - tblastx comparison file of *A. fumigatus* & *N. crassa*
6) A_fum_A_nid.comp - tblastx comparison file of *A. fumigatus* & *A. nidulans*
7) A_nid_P_anserina.comp - tblastx comparison file of *A. nidulans* & *P. anserina*
8) P_anserina_N_crassa.comp- tblastx comparison file of *P. anserina* & *N. crassa*.

First, open an ACT window and then open the annotation and the appropriate comparison files in the order of 1 – 5 – 2 – 6 – 3 – 7 – 4 – 8 – 1 (the numbers are designated above).

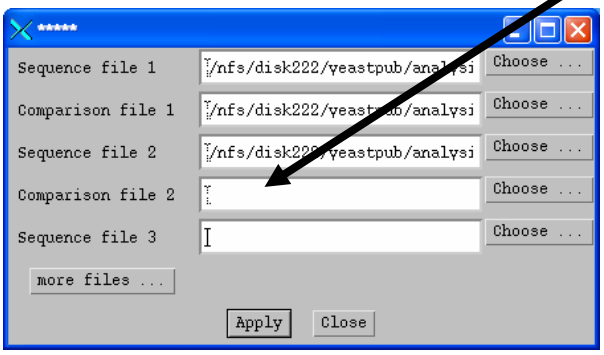
You will need to click on ‘more files’ to upload more than 2 sequences and the comparison flies.

Click on ‘apply’ after you have uploaded all the files.

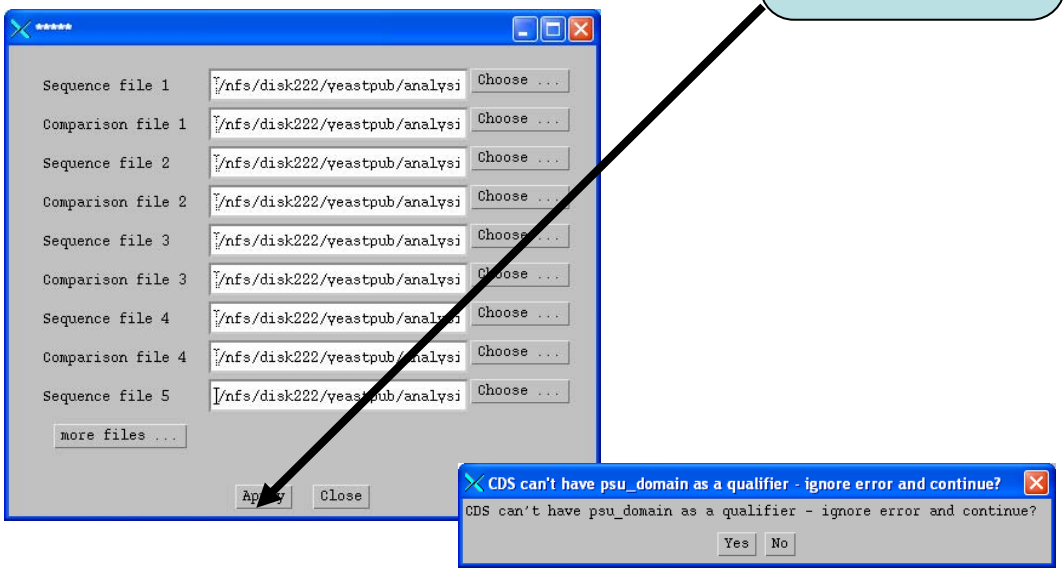
Upload the files in sequential order as described in the previous page



Click on here to load more files and select the appropriate file



Click on here to read all the files that you have selected.



Click on 'yes' if any small dialogue window appears while reading / opening the files.

Can you see any conserved gene order between the *A. fumigatus* & *A. nidulans* in the qut gene cluster?

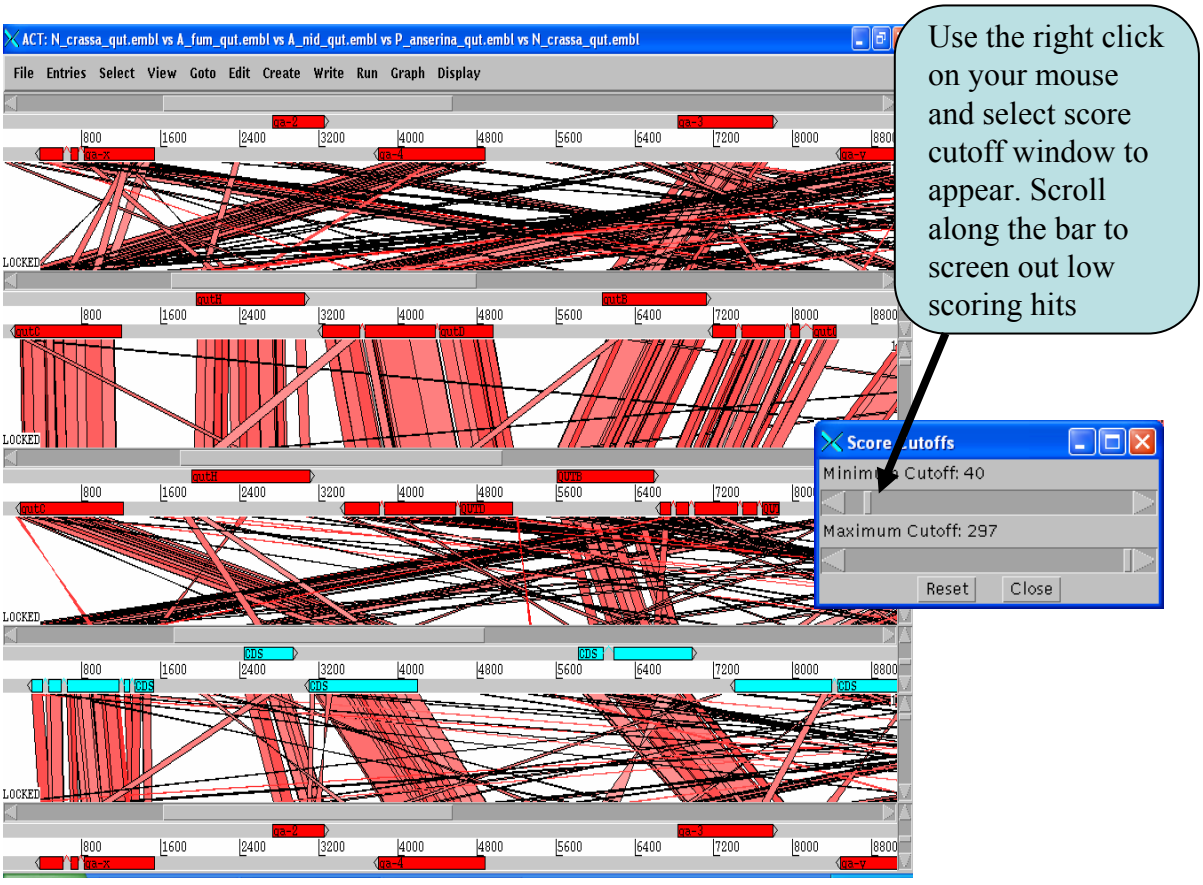
Can you obtain a clearer picture of the ACT 4-way comparison figure by filtering out the low scoring segments, using the blast score cut off feature which you have used previously.

Zoom in and look at some of the genes encoded within theses regions. View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.

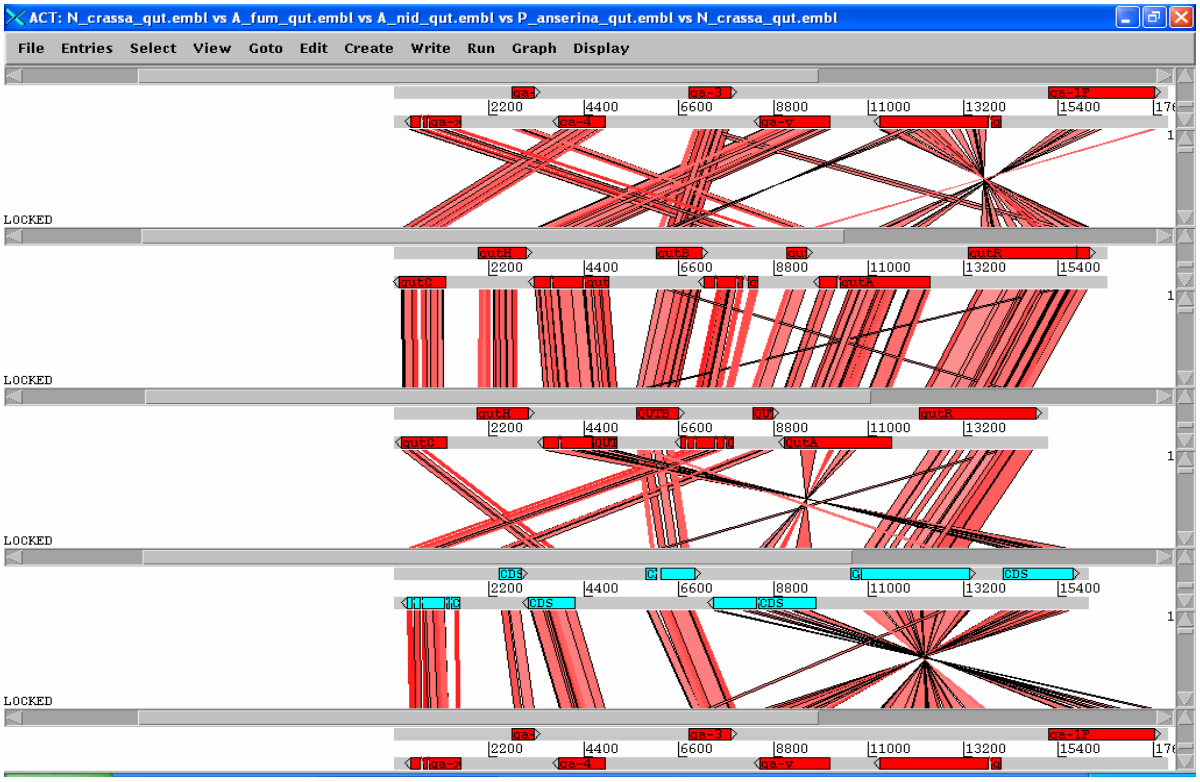
By comparing the blast similarity matches, assign your own annotation (gene product) to the predicted gene models (the blue genes) on the *P. anserina* gene model file.

Can you identify any gene NOT present in the qut cluster of ALL four fungi?

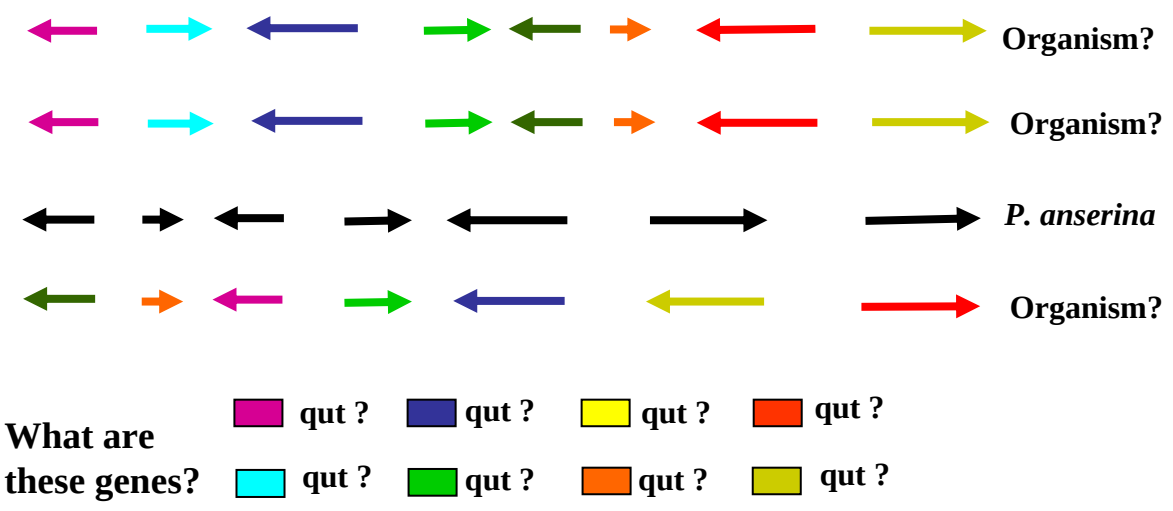
Note down the gene order (and direction of transcription) in each after you have completed annotation of the *P. anserina* gnes in the *qut* cluster.



After filtering out the low-scoring blast matches, you should be able to see a figure like the image below.



After comparing the arrangement of genes in the *qut* cluster in these fungi, do you agree with the schematic diagram (not in scale) below where each colour represents a specific type of gene in the quinic acid utilisation gene cluster and each set of clustered genes represents the *qut* cluster one of the organsims. Before you do this you need to annotate the *P. anserina* genes shown as black arrows.



Module 3
Generating ACT comparison files using BLAST

Introduction

In Module 2 you used ACT to visualize pairwise BlastN or TBlastX comparisons between DNA sequences. In order to use ACT to investigate your own sequences of interest you will have to generate your own pairwise comparison files. ACT is written so that it will read the output of several different comparison file formats; these are outlined in appendix II. Two of the formats can be generated using Blast software freely downloadable from the NCBI (appendix X). Both Windows and Linux versions of the software are available which can be loaded onto a PC or Mac.

For the purposes of this module the NCBI Blast distribution software has already been installed locally and therefore ready to use. To give you an idea of how easy it is to download and install the software on a PC we have included a step-by-step guide in the appendixes (Appendix X). The example shown in appendix X is for downloading onto a PC with Windows XP. The exercises in this module are based on the Linux version of the Blast software. Although the operating systems are different, the command lines used to run the programs are the same. One of the main differences between the two operating systems is that in Windows the Blast program command line is run in the DOS Command Prompt window, whereas in Linux it is run from a Xterminal window.

Aims
The aim of this module is to demonstrate how you can generate you own comparison files for ACT from a stand-alone version of the Blast software. In this module you will use Blast to generate comparison files for sequences that you have downloaded from the EBI genomes web resource. A copy of the Blast software has been installed locally. You will run Blast from the command-line using two different programs from the NCBI Blast distribution to generate ACT-readable comparison file for two small sequences (plasmids), and for two large sequences (whole genomes).

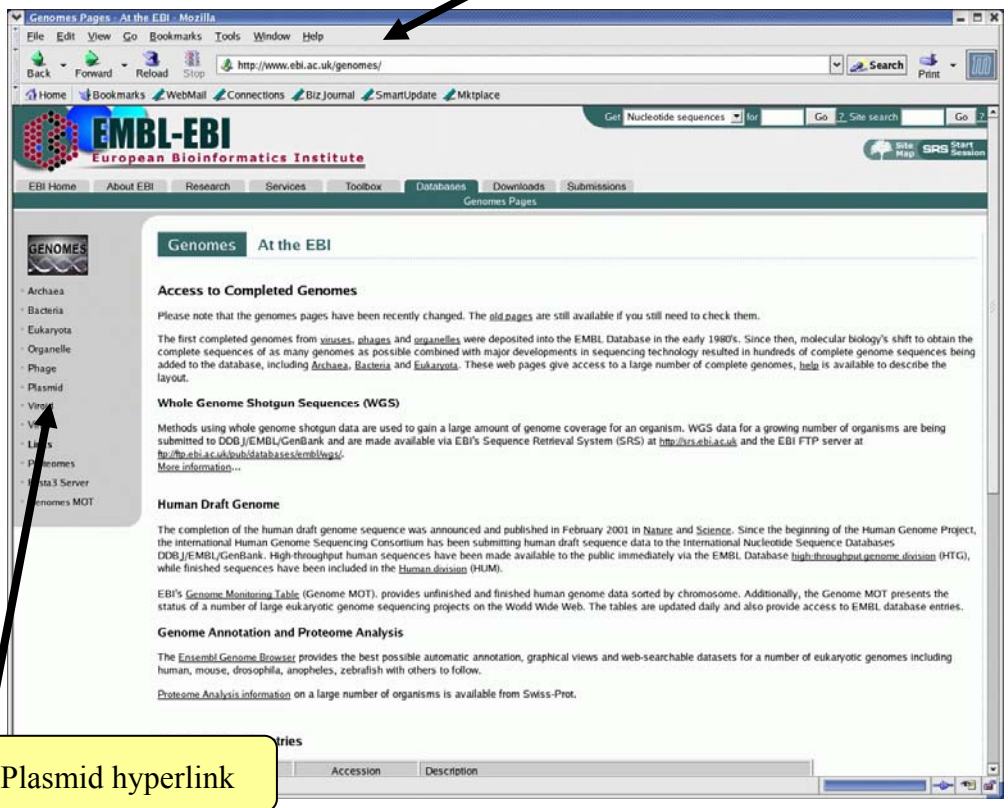
Exercise 1

In this exercise you are going to download two plasmid sequences in EMBL format from the EBI genomes web page. You are then going to use Artemis to write out the DNA sequences of both plasmids in FASTA format. These two FASTA format sequences will then be compared using BlastN to identify regions of DNA-DNA similarity and write out a ACT readable comparison file.

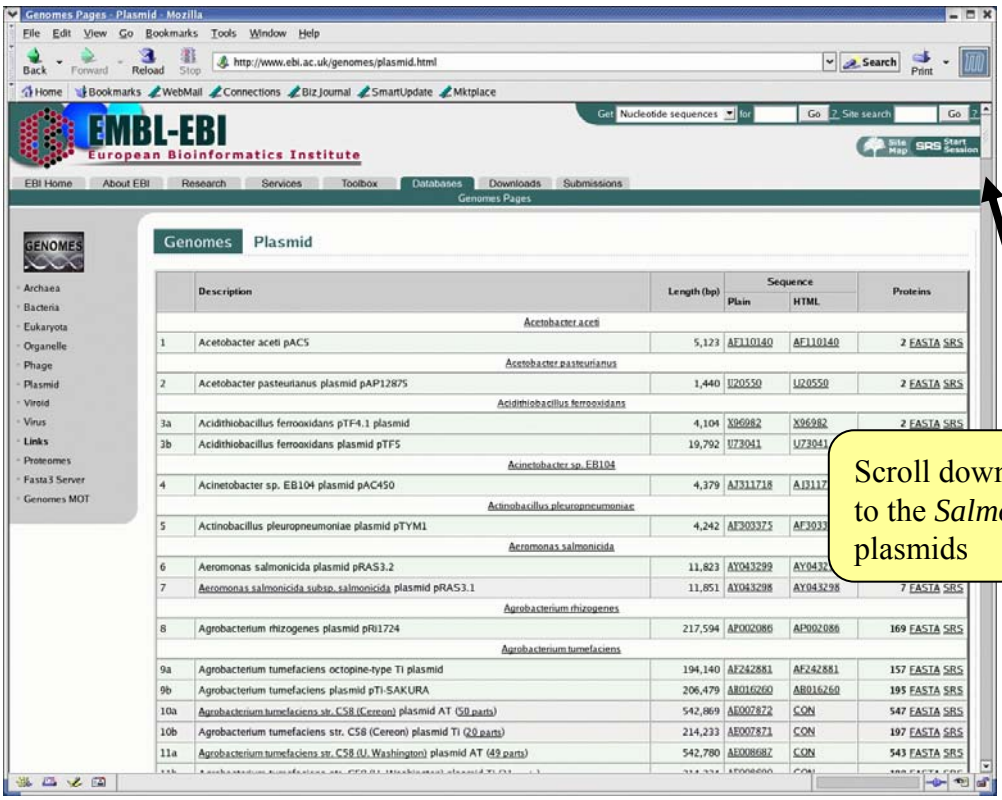
The plasmids chosen for this comparison are the multiple drug resistance incH1 plasmid pHCM1 from the sequenced strain of *Salmonella typhi* CT18 originally isolated in 1993, and R27, another incH1 plasmid first isolated from *S. typhi* in the 1960s.

Downloading the *S. typhi* plasmid sequences

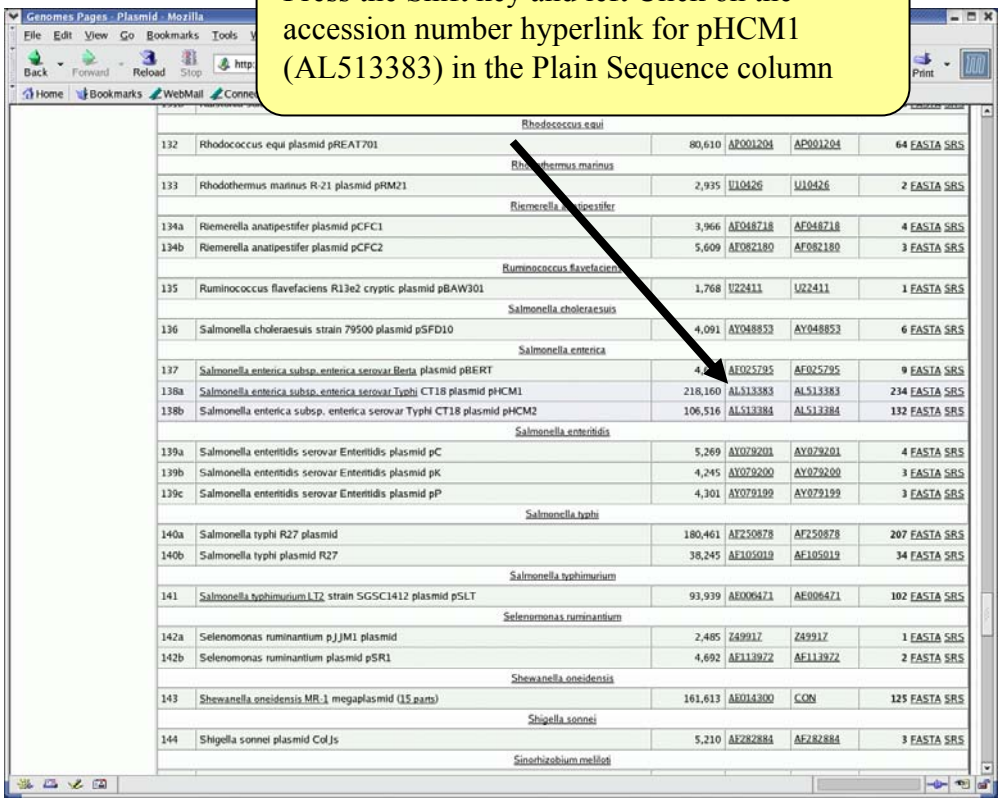
Go to the EBI genomes web page (http://www.ebi.ac.uk/genomes)

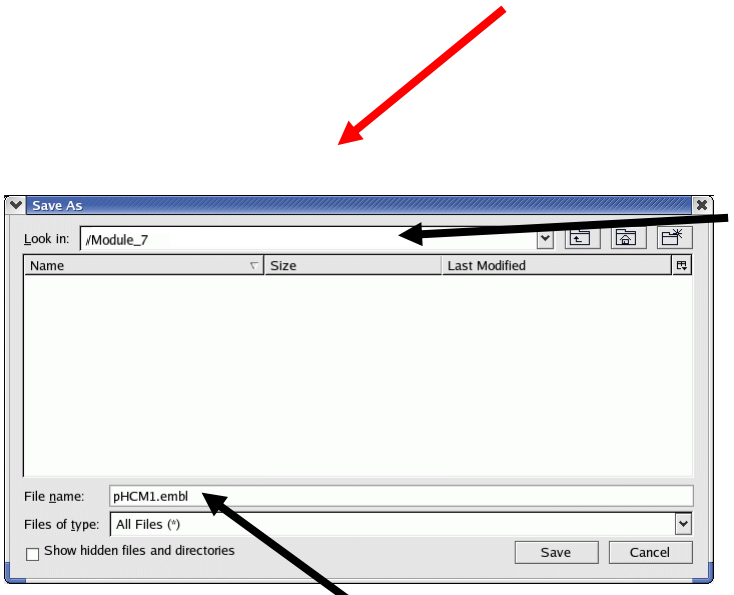


Click on the Plasmid hyperlink



Scroll down the page to the *Salmonella* plasmids

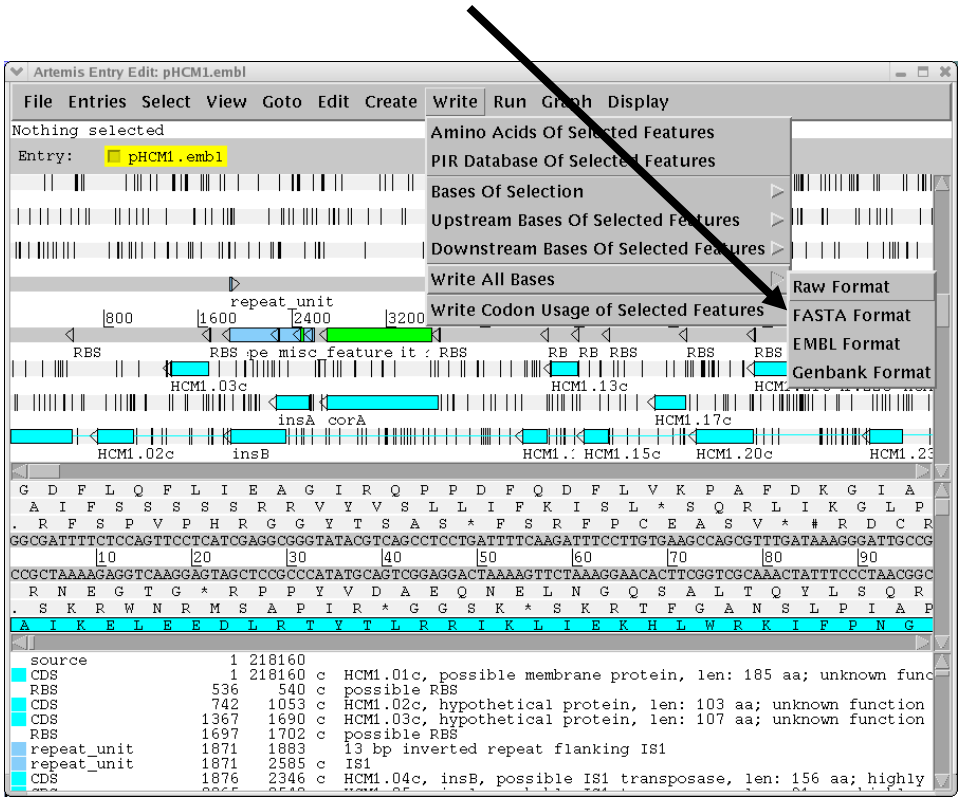


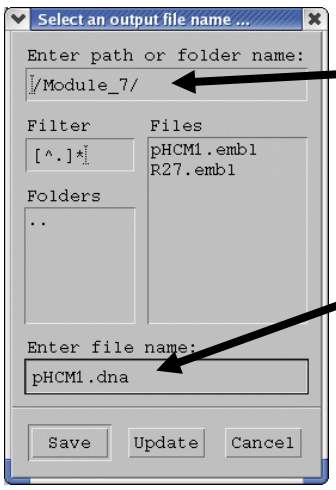


Repeat for the *Salmonella typhi* R27 plasmid (AF250878). Be careful when choosing the plasmid to download as there is also a *Salmonella typhi* plasmid R27 entry (AF105019), the one that you want is the larger of the two, 180,461 kb as opposed to 38,245 kb. Save as R27.embl.

In order to run BlastN you require two DNA sequences in FASTA format. The pHCM1 and R27 sequences previously downloaded from the EBI are EMBL format files, i.e. they contain protein coding information and the DNA sequence. In order to generate the DNA files in FASTA format, Artemis can be used as follows.

Load up the plasmid EMBL files in **Artemis** (each plasmid requires a separate Artemis window), select **Write**, **Write All Bases**, **FASTA format**.





Save the DNA sequence in the **Module_3** directory

Save as pHCM1.dna

Also do this for R27.embl

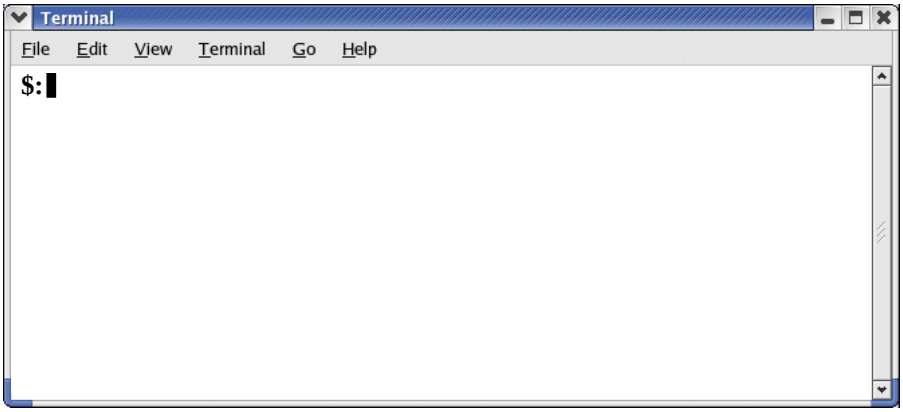
Running Blast

There are several programs in the Blast package that can be used for generating sequence comparison files. For a detailed description of the uses and options see the appropriate README file in the Blast software directory (see appendix X).

In order to generate comparison files that can be read into ACT you can use the **Blastall** program running either BlastN (DNA-DNA comparison) or TBlastX (translated DNA-translated DNA comparison) protocols.

As an example you will run a BlastN comparison on two relatively small sequences; the pHCM1 and R27 plasmids from *S. typhi*. In principle any DNA sequences in FASTA format can be used, although size becomes and issue when dealing with sequences such whole genomes of several Mb (see exercise 2 in this module). When obtaining nucleotide sequences from databases such as EMBL using a server such as SRS (http://srs.ebi.ac.uk), it is possible to specify that the sequences are in FASTA format.

To run the blast software you will need an Xterminal window like the one below. If you do not already have one opened, you can open a new window by clicking on the Xterminal icon on the menu bar at the bottom of your screen.



Make sure you are in the Module_3 directory. You should now see both the new FASTA files for the pHCM1 and R27 sequences in the Module_3 directory as well as their respective EMBL format files.
(Hint: You can use the **pwd** command to check the present working directory, the **cd** command to change directories, and the **ls** command will list the contents of the present working directory).

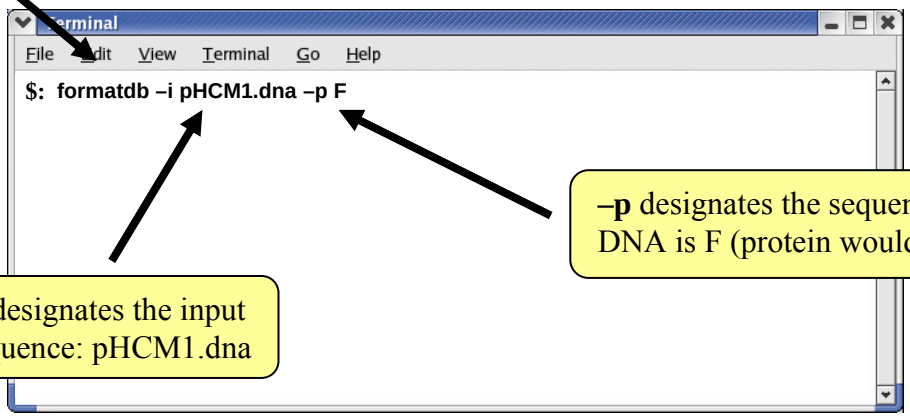
When comparing sequences in Blast, one sequence is designated as a **database** sequence, and the other the **query** sequence. Before you run Blast you have to format one of the sequences so that Blast recognises it as a database sequence. **formatdb** is a program that does this and comes as part of the NCBI Blast distribution.

You will treat pHCM1.dna as the **database** sequence and R27.dna as the **query** sequence

At the Command Prompt type:
formatdb –i pHCM1.dna –p F

Press **Return**

formatdb is the database format program



–p designates the sequence type: DNA is F (protein would be T)

–i designates the input sequence: pHCM1.dna

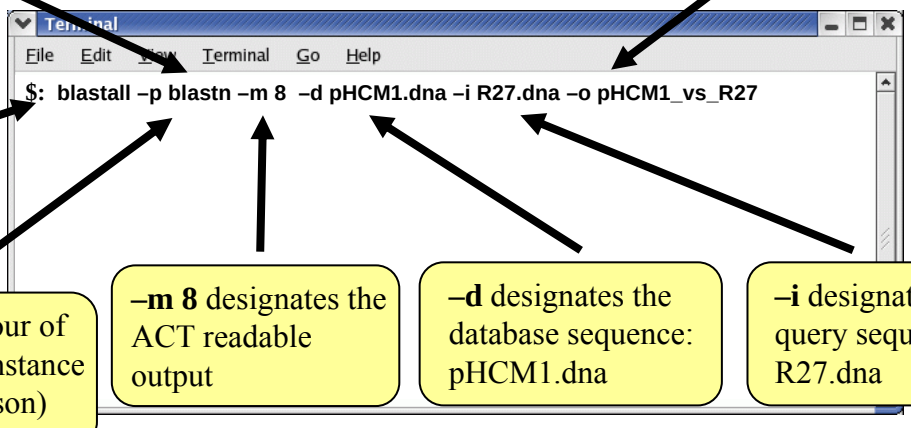
Now you can run the Blast on the two plasmid sequences. The program that you are going to use is **blastall**. In addition to the standard command line inputs we have to add an additional flag (**-m 8**) to the command line so that the Blast output can be read by ACT. This specifies that the output of Blast is in one line per entry format (see appendix II).

At the Command Prompt type:
blastall –p blastn –m 8 –d pHCM1.dna –i R27.dna –o pHCM1_vs_R27

Press **Return**

tblastx could be substituted here if a translated DNA-translated DNA comparison was required

–o designates the output file: pHCM1_vs_R27



blastall is the Blast program

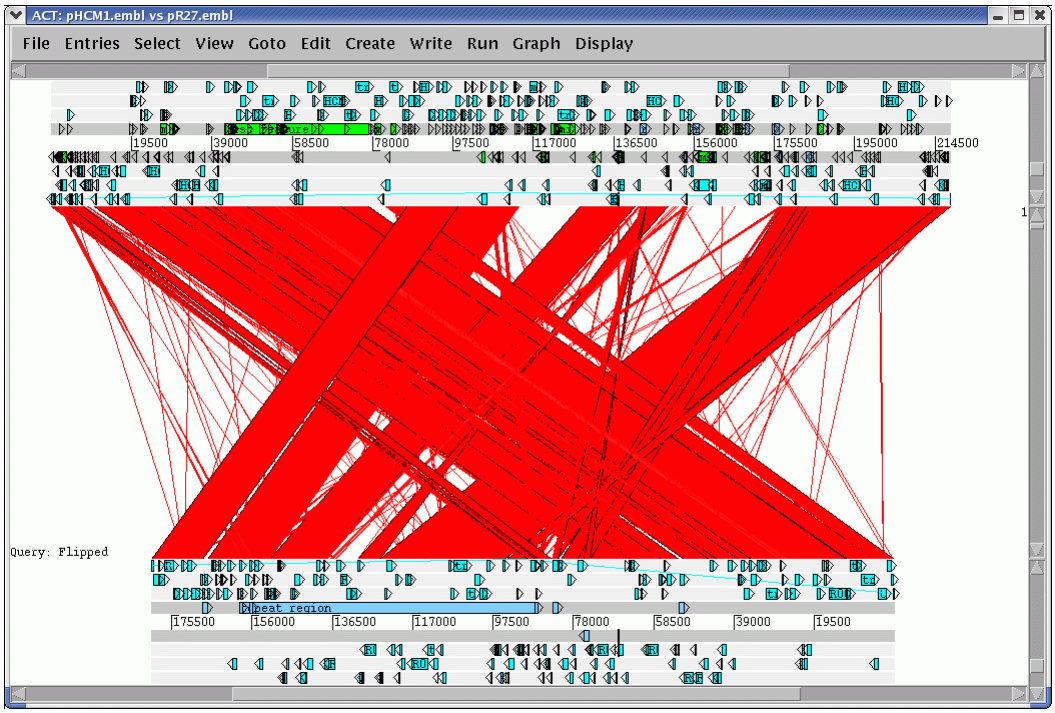
–p designates the flavour of Blast: **blastn** (in this instance a DNA-DNA comparison)

–m 8 designates the ACT readable output

–d designates the database sequence: pHCM1.dna

–i designates the query sequence: R27.dna

The pHCM1_vs_R27 comparison file can now be read into ACT along with the pHCM1.embl and R27.embl (or pHCM1.dna and R27.dna) sequence files.



The result of the BlastN comparison shows that there are regions of DNA shared between the plasmids; pHCM1 shares 169 kb of DNA at greater than 99% sequence identity with R27. Much of the additional DNA in the pHCM1 plasmid appears to have been inserted relative to R27 and encodes functions associated with drug resistance. What antibiotic resistance genes can you find in the pHCM1 plasmid that are not found in R27?

The two plasmids were isolated more than 20 years apart. The comparison suggest that there have been several independent acquisition events that are responsible for the multiple drug resistance seen in the more modern *S. typhi* plasmid.

Exercise 2

In the previous exercise you used BlastN to generate a comparison file for two relatively small sequences (>500,000 kb). In the next exercise we are going to use another program from NCBI Blast distribution, **megablast**, that can be used for nucleotide sequence alignment searches, i.e. DNA-DNA comparisons. If you are comparing large sequences such as whole genomes of several Mb, the **blastall** program is not suitable. The Blast algorithms will struggle with large DNA sequences and therefore the processing time to generate a comparison file will increase dramatically.

Megablast uses a different algorithm to Blast which is not as stringent which therefore makes the program faster. This means that it is possible to generate comparison files for genome sequences in a matter of seconds rather than minutes and hours.

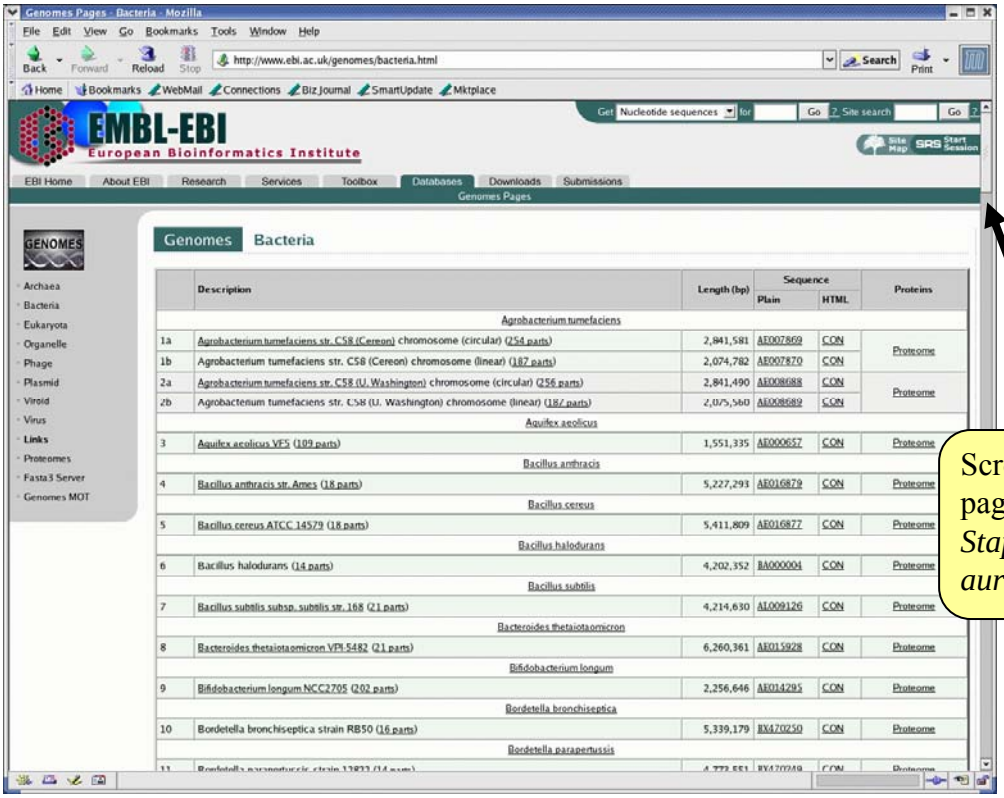
There are some drawbacks to using this program. Firstly, only DNA-DNA alignments (BlastN) can be performed using **megablast**, rather than translated DNA-DNA alignments (TBlastX) as can be using **blastall**. Secondly as the algorithm used is not as stringent, megablast is suited to comparing sequences with high levels of similarity such as genomes from the same or very closely related species.

In this exercise you are going to download two *Staphylococcus aureus* genome sequences from the EBI genomes web page and use Artemis to write out the FASTA format DNA sequences for both as before in exercise 1. These two FASTA format sequences will then be compared using **megablast** to identify regions of DNA-DNA similarity and write out an ACT readable comparison file.

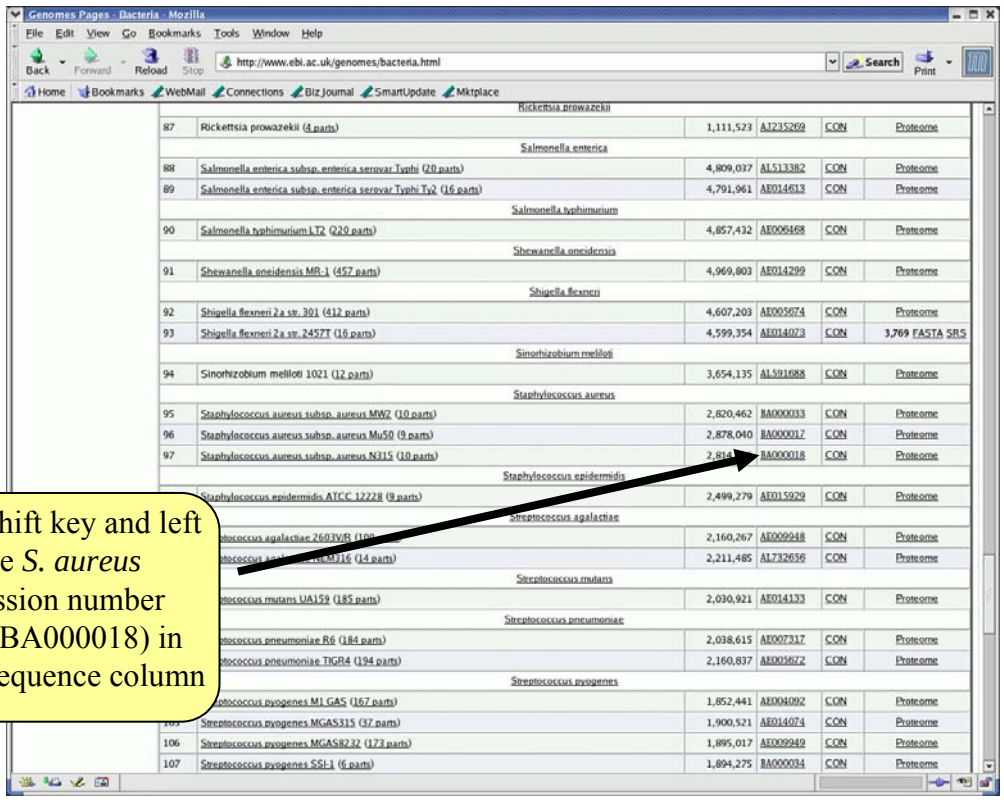
The genomes that have been chosen for this comparison are from a hospital-acquired methicillin resistant *S. aureus* (MRSA) strain N315 (BA000018), and a community-acquired MRSA strain MW2 (BA000033).

Downloading the *S. aureus* genomic sequences

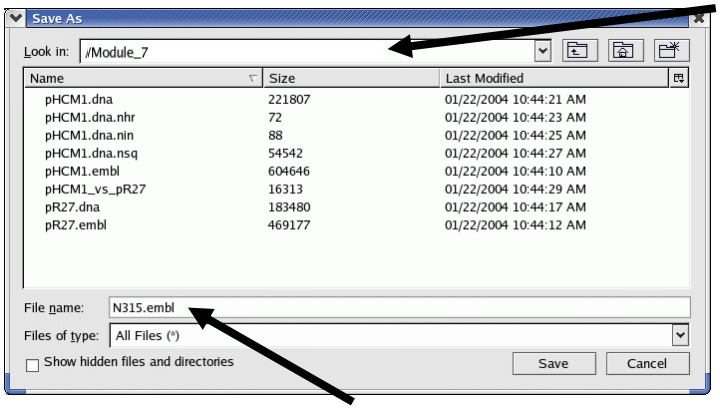
Go to the EBI genomes web page (http://www.ebi.ac.uk/genomes) as before in exercise 1, and click on the **Bacteria** hyperlink



Scroll down the page to the *Staphylococcus aureus* genomes



Press the Shift key and left Click on the *S. aureus* N315 accession number hyperlink (BA000018) in the Plain Sequence column



Save the EMBL sequence in the **Module_3** directory

Save the file as N315.embl

Repeat for the *S. aureus* MW2 genome (BA000033). Be careful when choosing the genome to download as there is another *S. aureus* genome entry for strain Mu50 (BA000017). Save as MW2.embl.

Generate DNA files in FASTA format using Artemis for both the genome sequences as previously done in exercise 1.

(Hint: In **Artemis** (each genome requires a separate Artemis window), select **Write**, **Write All Bases**, **FASTA format**).

Save the DNA sequences as N315.dna and MW2.dna for the respective genomes.

Running Blast

In the previous exercise you used the **blastall** program to run BlastN on two plasmid sequences. As the genome sequences are larger (~2.8 Mb) you are going to run **megablast**, another program from the NCBI Blast distribution that can generate comparison files in a format that ACT can read (see appendix II). For a detailed description of the uses and options in **megablast** see the megablast README file in the Blast software directory (appendix X).

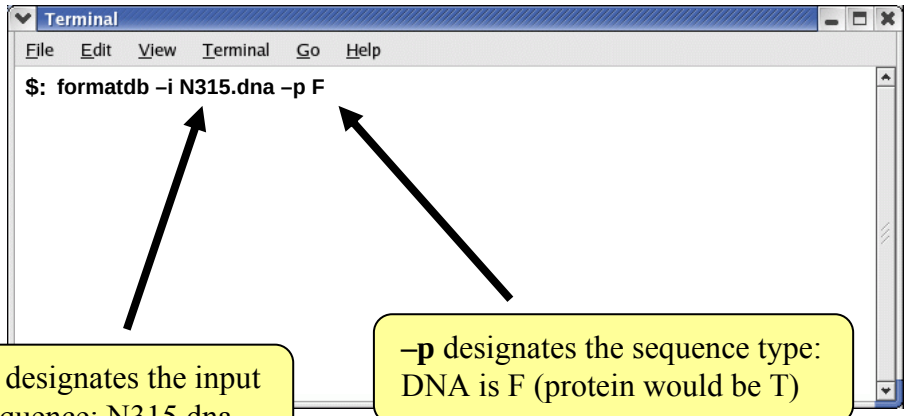
As before you will run the program from the command line in an Xterminal window.

Like Blast, **megablast** requires that one sequence is designated as a **database** sequence and the other the **query** sequence. Therefore one of the sequences has to be formatted so that Blast recognises it as a database sequence. This can be done as before using **formatdb.**

We will treat N315.dna as the **database** sequence and MW2.dna as the **query** sequence

At the Command Prompt type:
formatdb –i N315.dna –p F

Press **Return**



–i designates the input sequence: N315.dna

–p designates the sequence type: DNA is F (protein would be T)

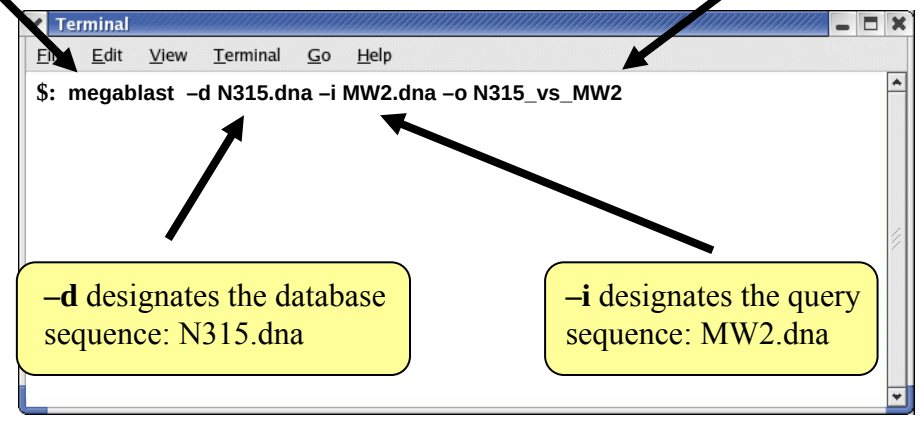
Now we can run the **megablast** on the two MRSA genome sequences. The default output format is one line per entry that ACT can read, therefore there is no need to add an additional flag to the command line (see appendix II).

At the Command Prompt type:
megablast –d N315.dna –i MW2.dna –o N315_vs_MW2

Press **Return**

megablast is the program

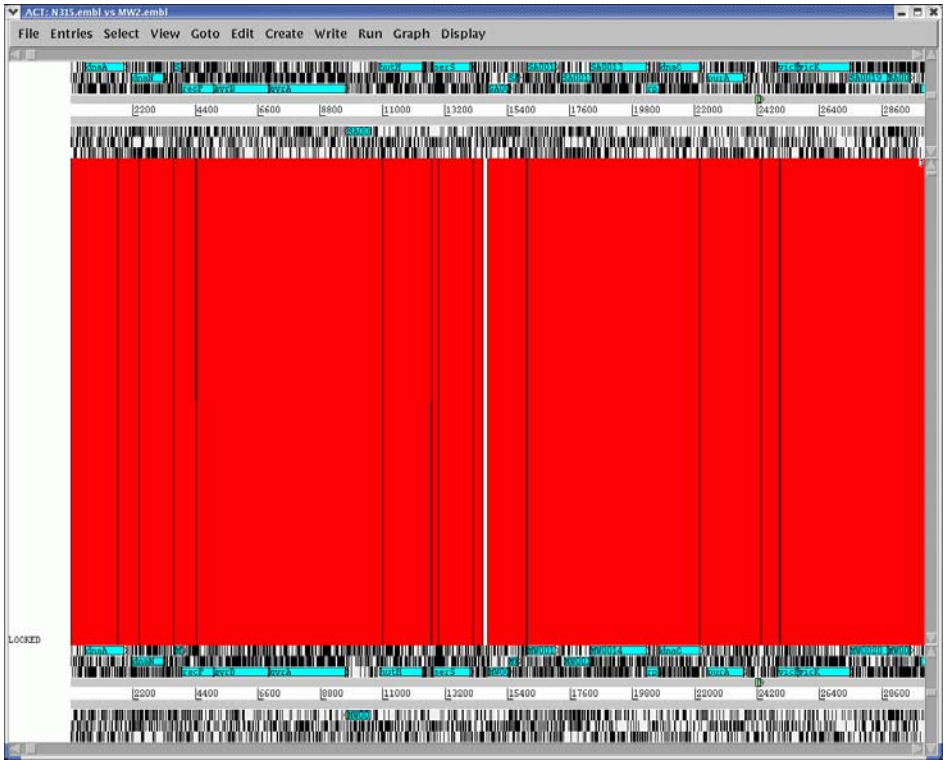
–o designates the output file: N315_vs_MW2



–d designates the database sequence: N315.dna

–i designates the query sequence: MW2.dna

The N315_vs_MW2 comparison file can now be read into ACT along with the N315.embl and MW2.embl (or N315.dna and MW2.dna) sequence files.



A comparison of the N315 and MW2 genomes in ACT using the **megablast** comparison reveals a high level of synteny (conserved gene order). This is perhaps not unsurprising as both genomes belong to strains of the same species. Using results of comparisons like these it is possible to identify genomic differences that may contribute to the biology of the bacteria and also investigate mechanisms of evolution.

Both N315 and MW2 are MRSA, however N315 is associated with disease in hospitals, and MW2 causes disease in the community and is more invasive. Scroll rightward in both genomes to find the first large region of difference. Examine the annotation for the genes in these regions. What are the encoded functions associated with these regions? What significance does this have for the evolution of methicillin resistance in these two *S. aureus* strains from clinically distinct origins?

Module 4
Jemboss

Module 5
Internet Genome Resources

Introduction

The preceding modules are concerned with predicting genes, and then trying to evaluate what they do. This module will deal firstly with some of the main ways that gene products are described using controlled vocabularies and secondly how you can use these description to quickly access genes from databases.
The module is split into three sections:

Section 1

EC numbers
- a very widely used system for describing enzymes. EC numbers can be used to find out additional information for an enzyme, such as possible orthologues, the biochemical pathway that it's involved in etc, or can be used to identify new enzymes.

Section 2

Gene Ontology
- a way to find genes based on descriptions of the molecular function, biological process or cellular component of their products.

Section 3

InterPro & UniProt
- An integrated documentation resource for protein families, domains and sites

Aims

The aim of this module will be to explore these controlled vocabularies using a series of worked examples.

Section 1

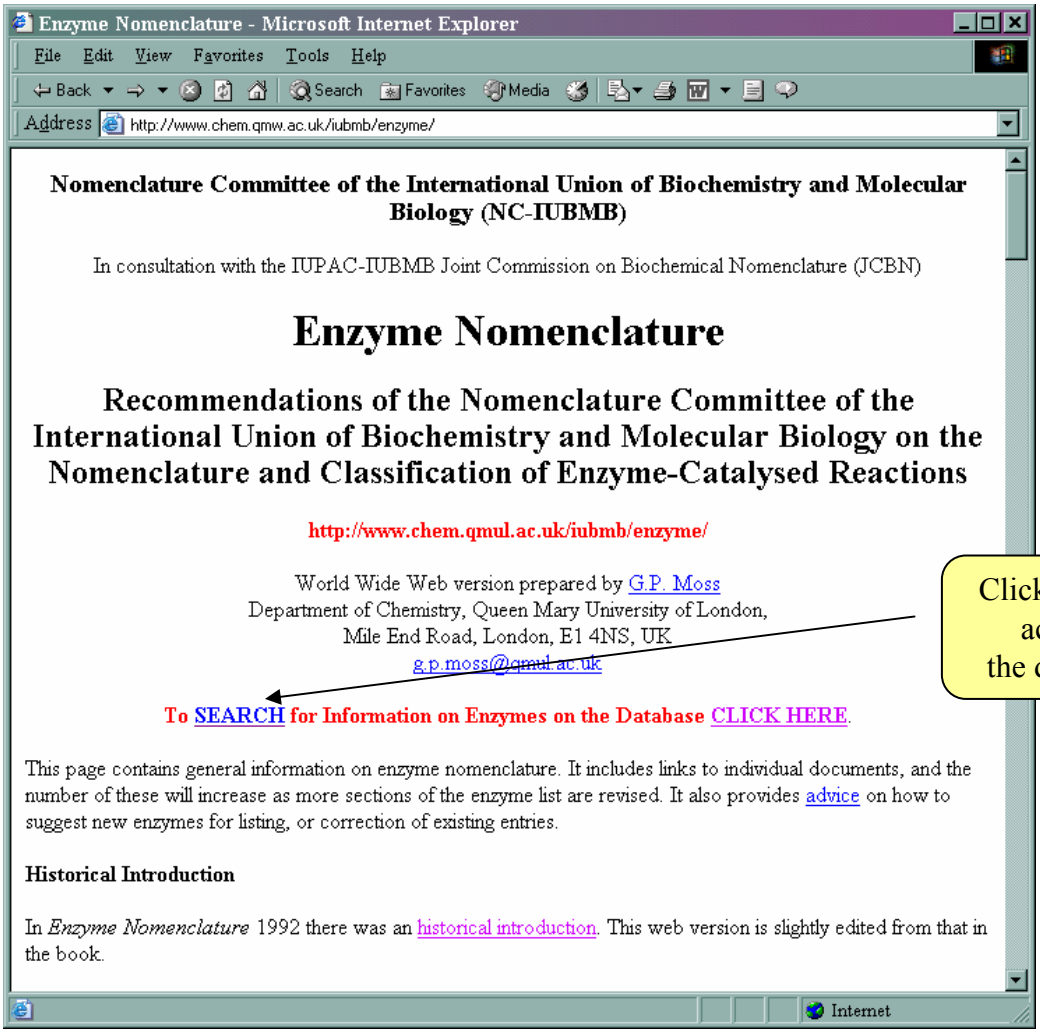
Exercise 1 Part I

1. What do *equilase* and 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase do? What kinds of pathways are they involved in?

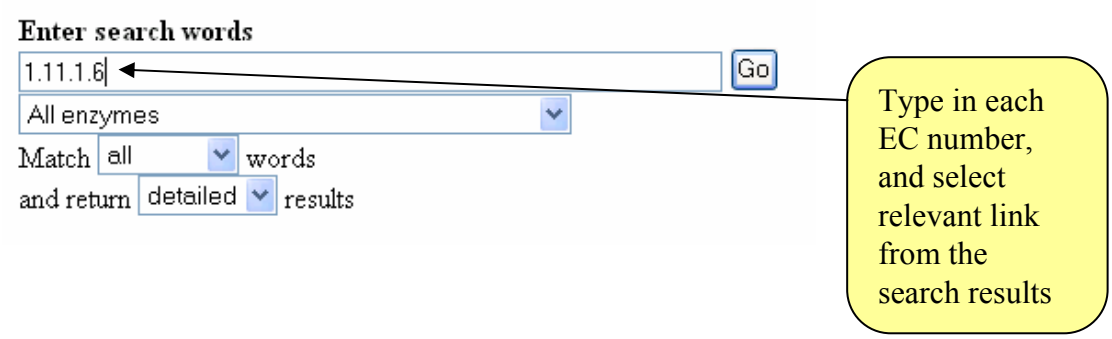
You probably won't have a very clear idea of what these enzymes are (even if you're a biochemist).

Use their EC numbers (EC 1.11.1.6 and EC 2.7.7.60, respectively) to find out more from the "official" Enzyme Nomenclature website

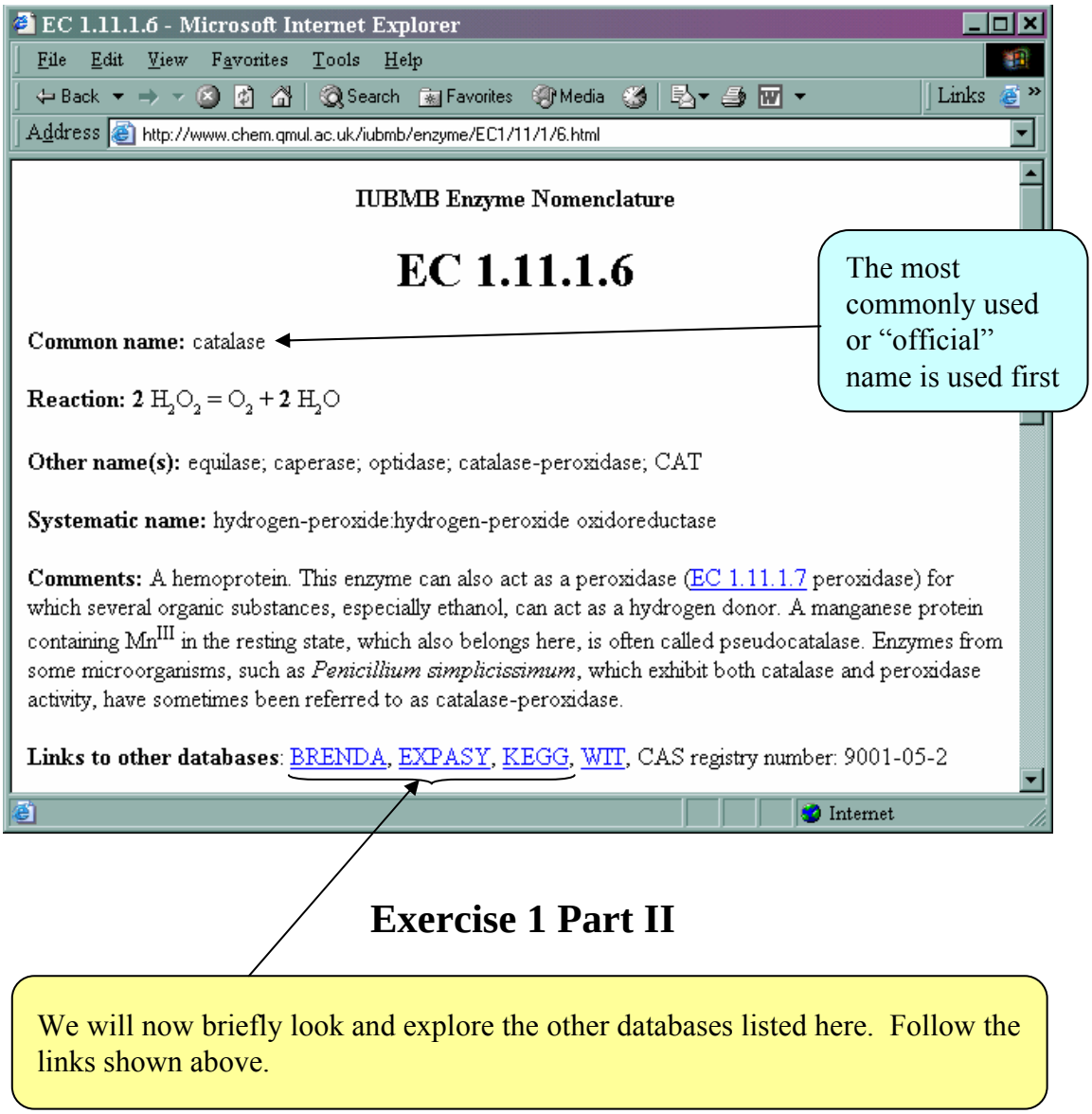
Go to this web address http://www.chem.qmw.ac.uk/iubmb/enzyme/ in your web browser window



Click here to access the database



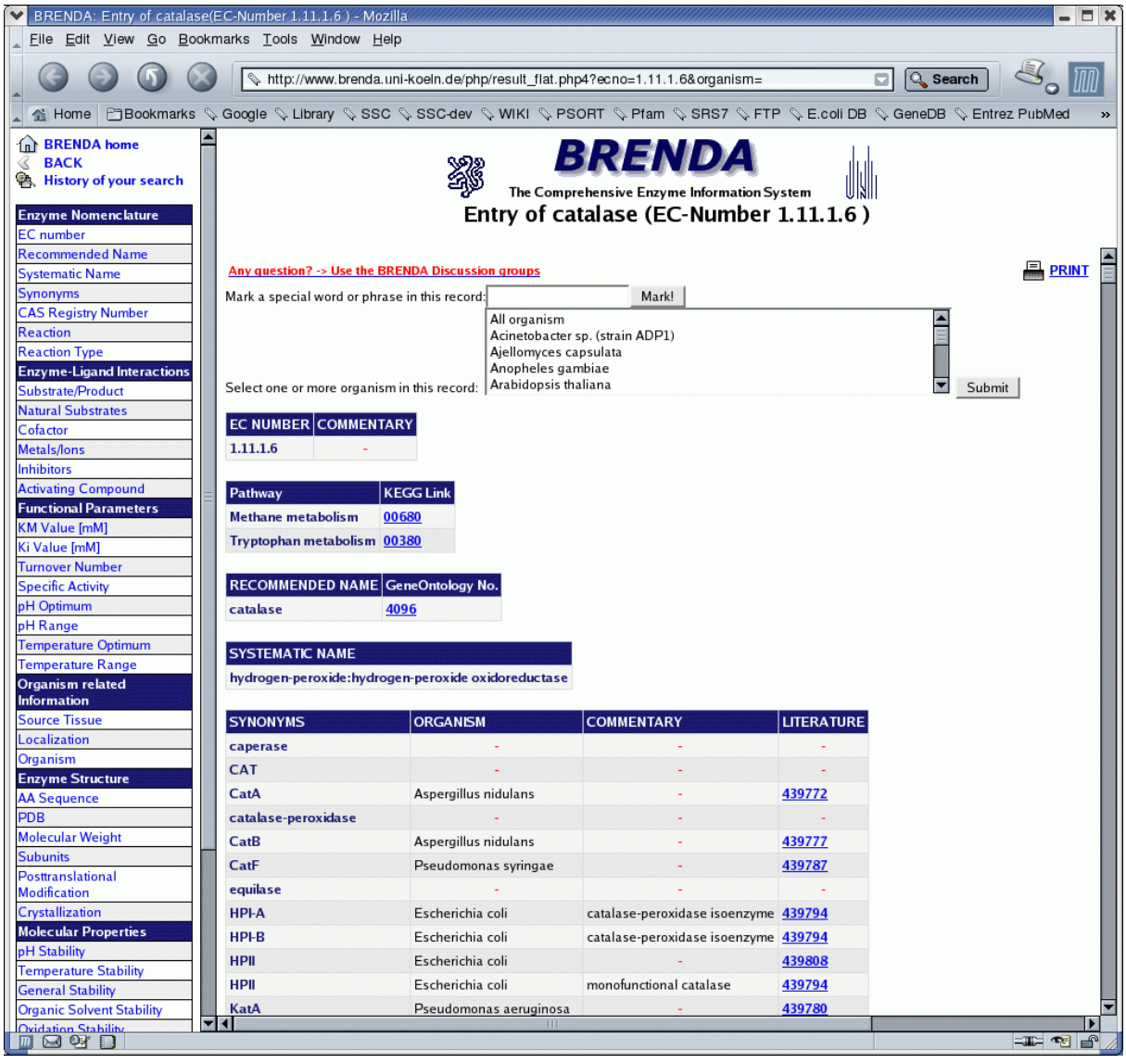
Each enzyme is represented by a separate web page in IUBMB.



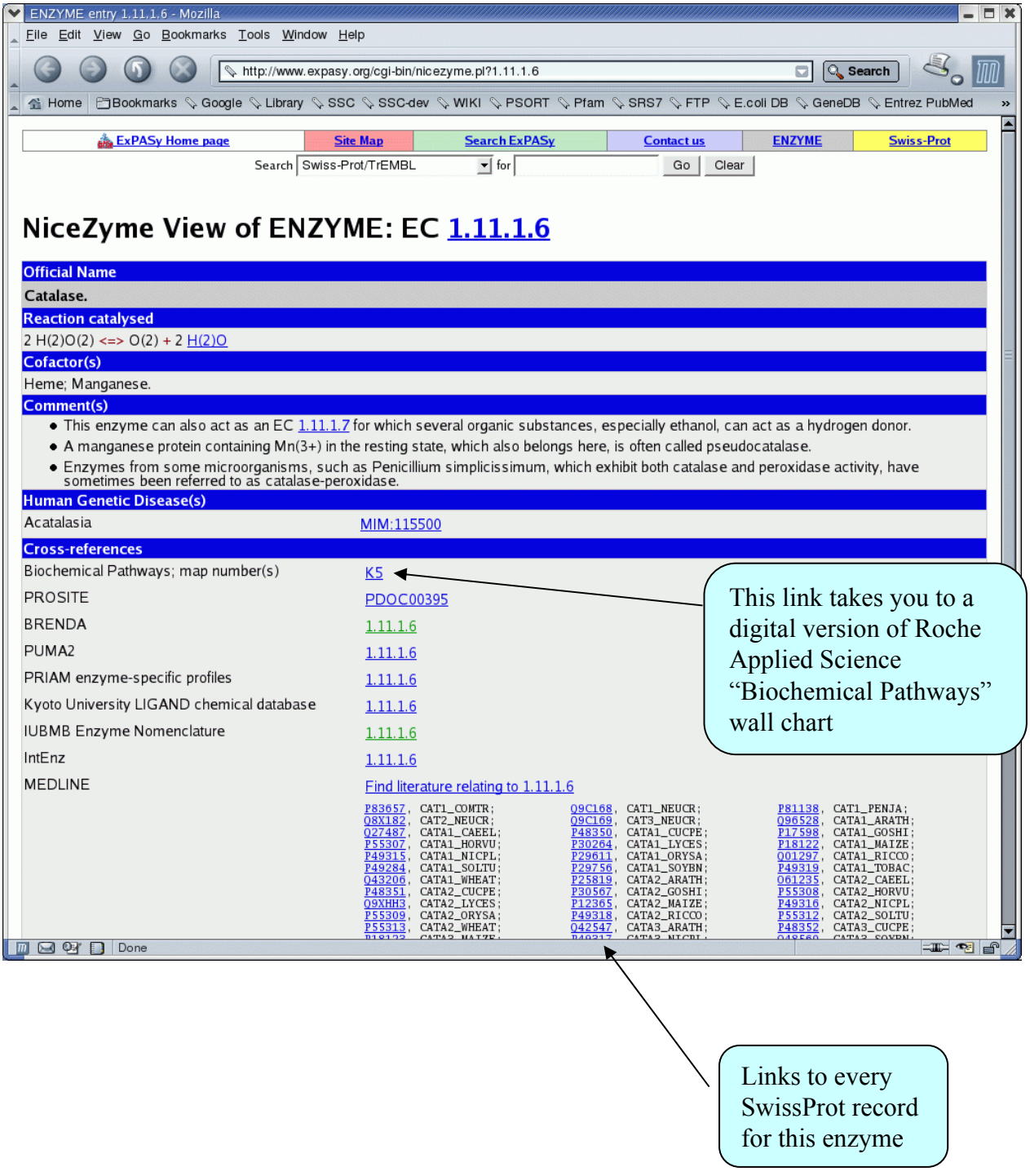
Exercise 1 Part II

We will now briefly look and explore the other databases listed here. Follow the links shown above.

The BRENDA database contains similar information to the IUBMB site…



EXPASY Database.view



KEGG view of EC 1.11.1.6

```
ENTRY       EC 1.11.1.6
NAME        catalase
            equilase
            caperase
            optidase
            catalase-peroxidase
            CAT
CLASS       Oxidoreductases
            Acting on a peroxide as acceptor
            Peroxidase
SYSNAME     hydrogen-peroxide:hydrogen-peroxide oxidoreductase
REACTION    2 H2O2 = O2 + 2 H2O
SUBSTRATE   H2O2
PRODUCT     H2O
            O2
COFACTOR    Heme
            Manganese
PATHWAY     PATH: MAP00380  Tryptophan metabolism
            PATH: MAP00680  Methane metabolism
GENES       HSA: 847(CAT)
            MMU: 12359(Cas1)
            RNO: 24248(Cat)
            DME: CG6871(Cat) CG9314
            CEL: Y54G11A.6
```

The KEGG database contains tools for analysing the enzymes in pathways

The use of the pathway maps at KEGG will be explored more fully in a later exercise in this module.

Exercise 1 Part III

Take your web browser back to the IUBMB search page and search using EC2.7.7.60 as before. For some enzymes, you can also get pathway information from their IUBMB pages e.g.

IUBMB Enzyme Nomenclature

EC 2.7.7.60

Common name: 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase

Reaction: CTP + 2-*C*-methyl-D-erythritol 4-phosphate = diphosphate + 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol

For diagram click here.

Click here to get a pathway diagram

Other name(s): MEP cytidylyltransferase

Systematic name: CTP:2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase

Comments: The enzyme from *Escherichia coli* requires Mg^{2+} or Mn^{2+}. ATP or UTP can replace CTP, but both are less effective. GTP and TTP are not substrates. Forms part of an alternative nonmevalonate pathway for terpenoid biosynthesis (for diagram, click here).

Links to other databases: BRENDA, EXPASY, KEGG, WIT, CAS registry number:

References:

1. Rohdich, F., Wungsintaweekul, J., Fellermeier, M., Sagner, S., Herz, S., Kis, K., Eisenreich, W., Bacher, A. and Zenk, M.H. Cytidine 5'-triphosphate-dependent biosynthesis of isoprenoids: YgbP protein of *Escherichia coli* catalyzes the formation of 4-diphosphocytidyl-2-*C*-methyl-D-erithritol. *Proc. Natl Acad. Sci. USA* 96 (1999) 11758-11763. [Medline UI: 10518523]

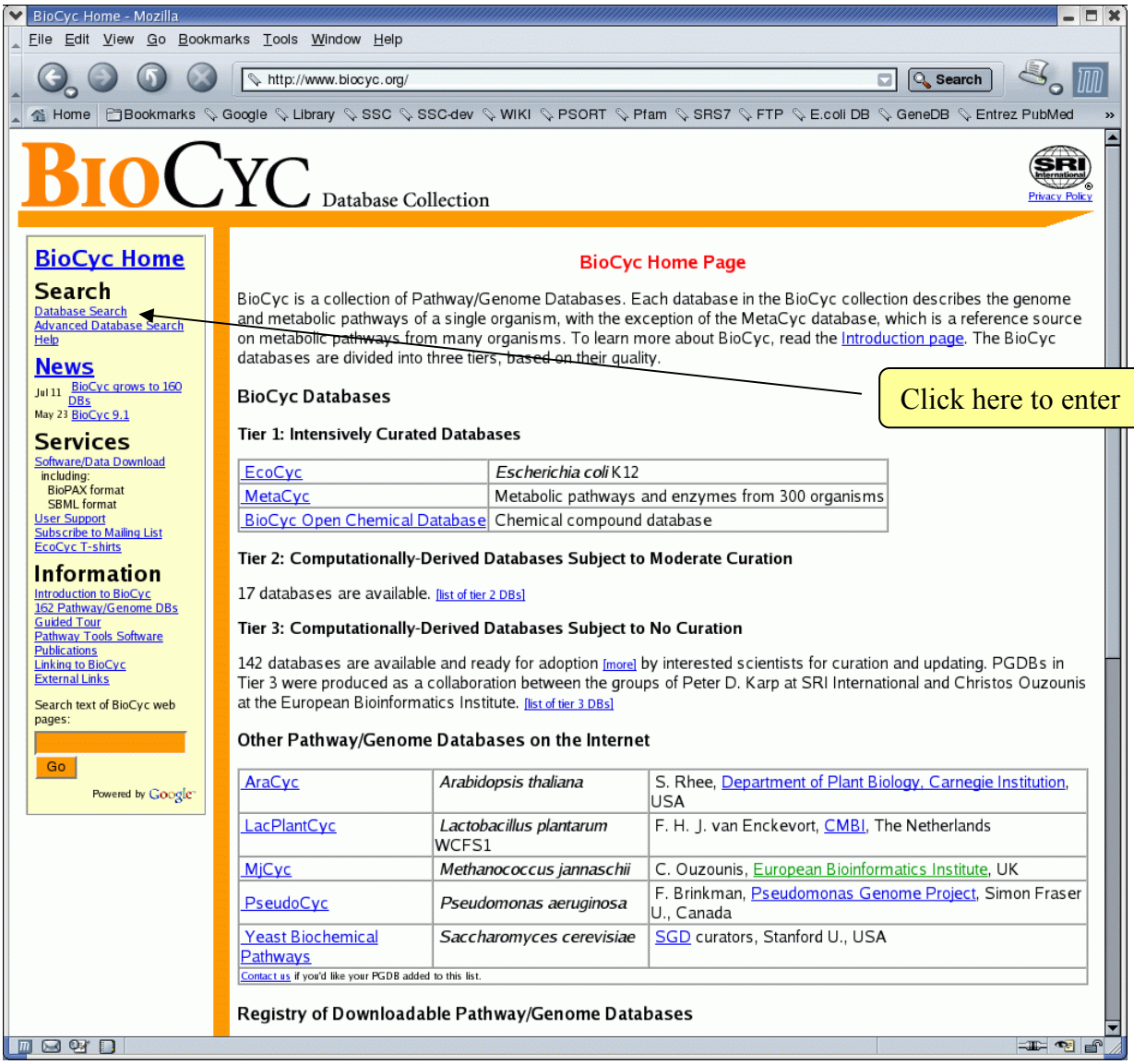
2. Kuzuyama, T., Takagi, M., Kaneda, K., Dairi, T. and Seto, H. Formation of 4-(cytidine 5'-diphospho)-2-*C*-methyl-D-erythritol from 2-*C*-methyl-D-erythritol 4-phosphate by 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase, a new enzyme in the nonmevalonate pathway. *Tetrahedron Lett.* 41 (2000) 703-706.

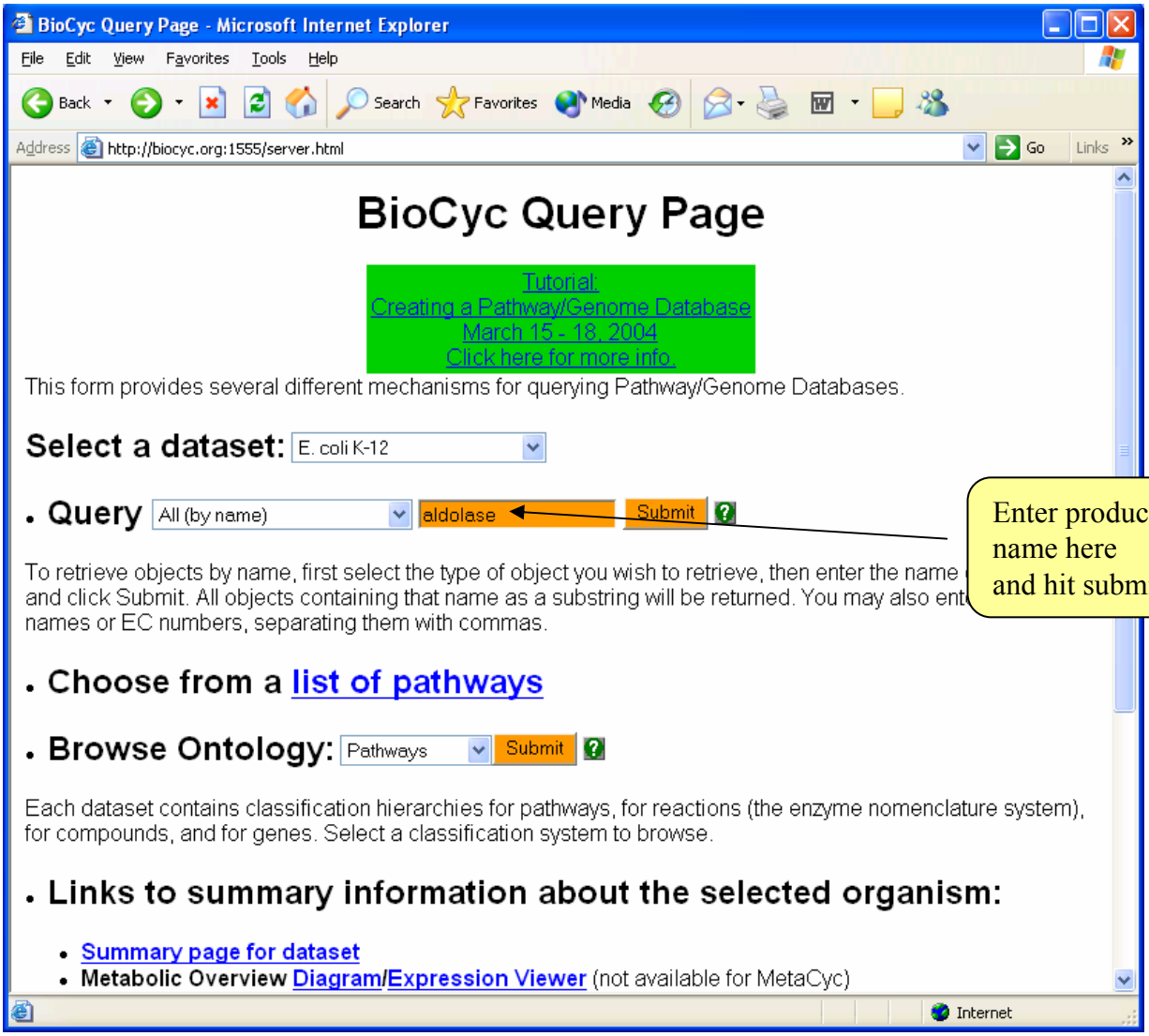
Exercise 2

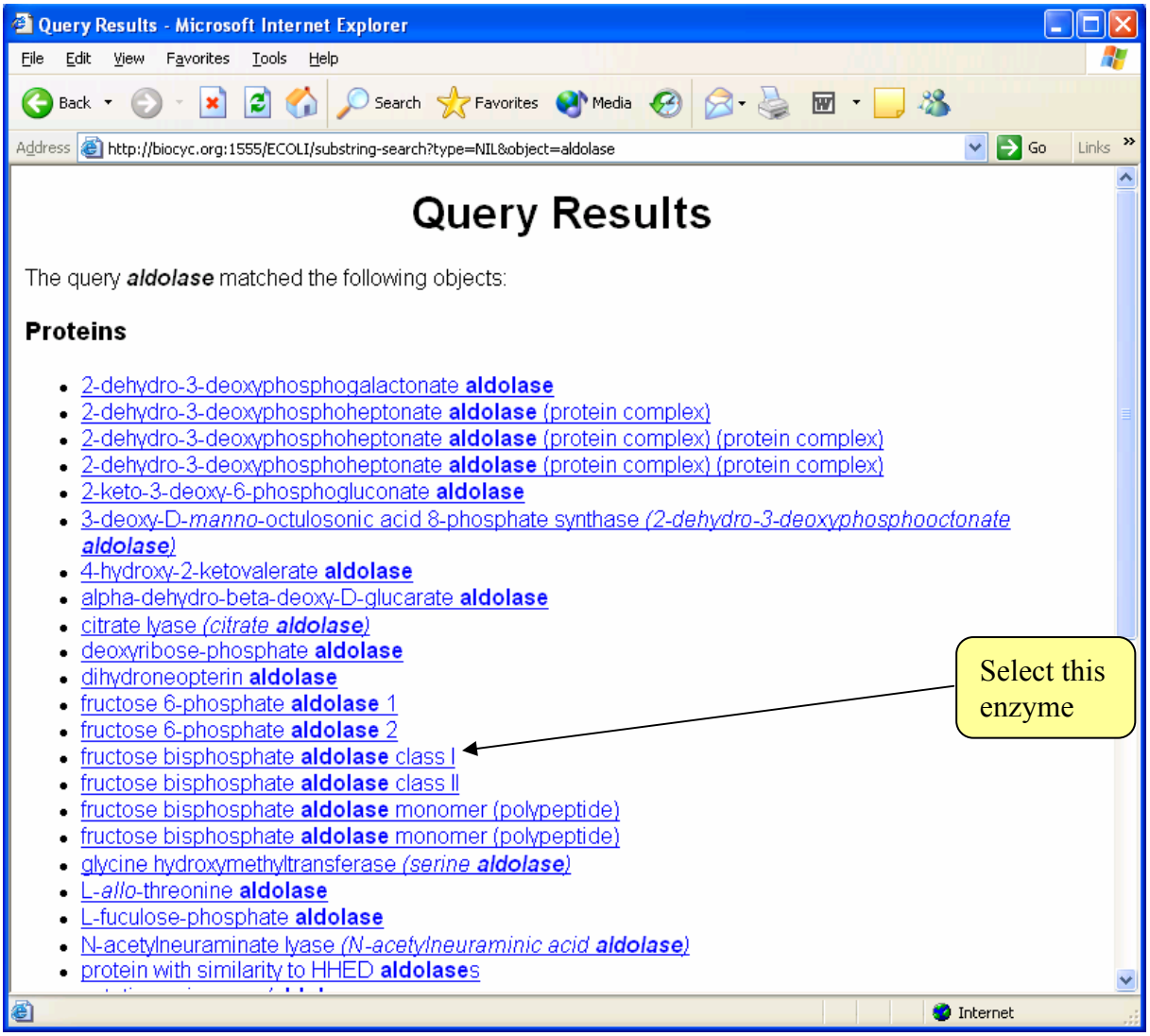
2. What is the EC number of Fructose-1,6-Bisphosphate aldolase

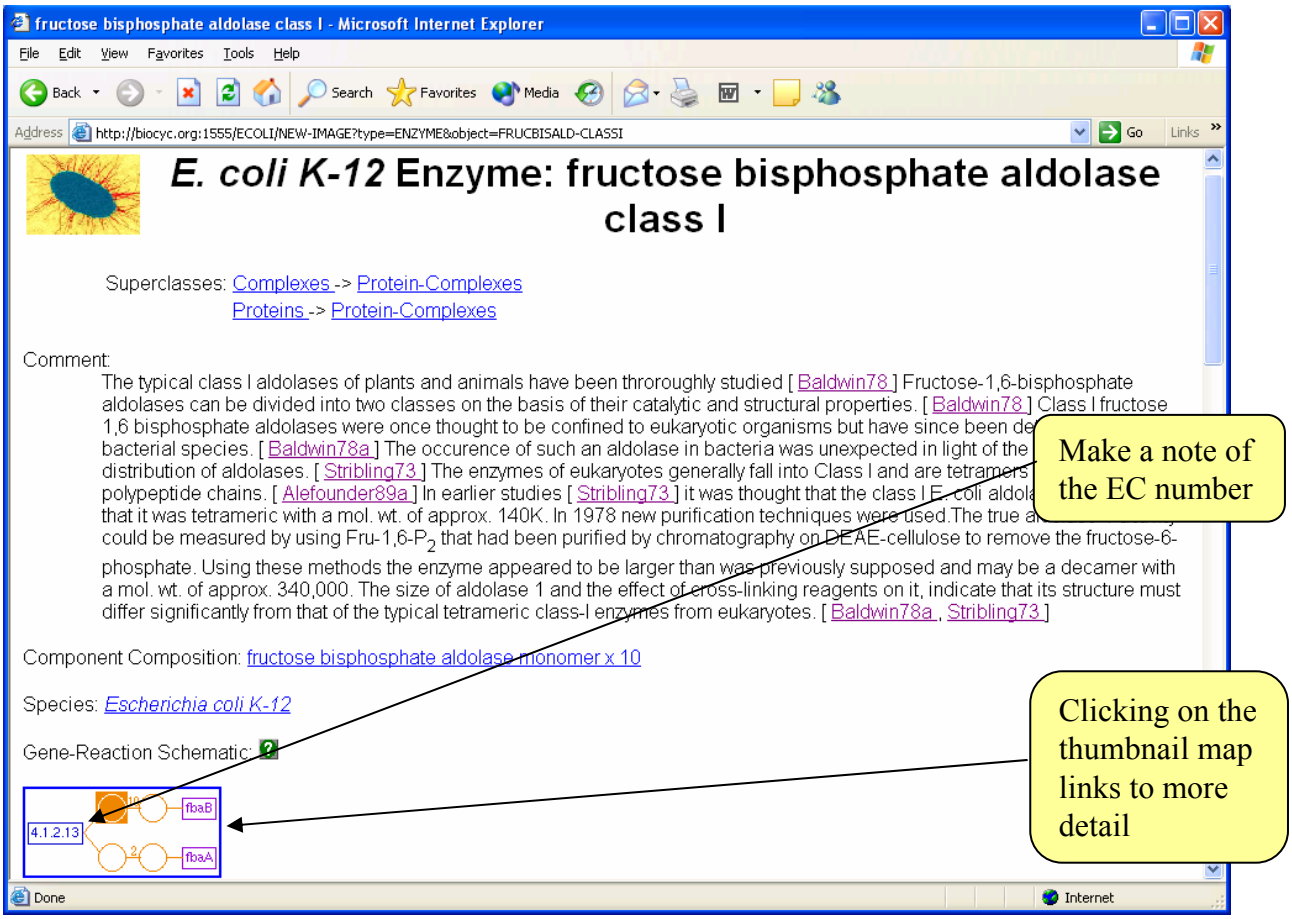
Use the BioCyc Genome Pathway Database to find out the EC number.

http://www.biocyc.org




BioCyc Query Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://biocyc.org:1555/server.html
Go Links
BioCyc Query Page
Tutorial:
Creating a Pathway/Genome Database
March 15 - 18, 2004
Click here for more info.
This form provides several different mechanisms for querying Pathway/Genome Databases.
Select a dataset: E. coli K-12
Query All (by name) aldolase Submit
To retrieve objects by name, first select the type of object you wish to retrieve, then enter the name
and click Submit. All objects containing that name as a substring will be returned. You may also ent
names or EC numbers, separating them with commas.
Choose from a list of pathways
Browse Ontology: Pathways Submit
Each dataset contains classification hierarchies for pathways, for reactions (the enzyme nomenclature system),
for compounds, and for genes. Select a classification system to browse.
Links to summary information about the selected organism:
Summary page for dataset
Metabolic Overview Diagram/Expression Viewer (not available for MetaCyc)
Internet
Enter product name here and hit submit


Query Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://biocyc.org:1555/ECOLI/substring-search?type=NIL&object=aldolase
Go Links
Query Results
The query aldolase matched the following objects:
Proteins
2-dehydro-3-deoxyphosphogalactonate aldolase
2-dehydro-3-deoxyphosphoheptonate aldolase (protein complex)
2-dehydro-3-deoxyphosphoheptonate aldolase (protein complex) (protein complex)
2-dehydro-3-deoxyphosphoheptonate aldolase (protein complex) (protein complex)
2-keto-3-deoxy-6-phosphogluconate aldolase
3-deoxy-D-manno-octulosonic acid 8-phosphate synthase (2-dehydro-3-deoxyphosphooctonate aldolase)
4-hydroxy-2-ketovalerate aldolase
alpha-dehydro-beta-deoxy-D-glucarate aldolase
citrate lyase (citrate aldolase)
deoxyribose-phosphate aldolase
dihydroneopterin aldolase
fructose 6-phosphate aldolase 1
fructose 6-phosphate aldolase 2
fructose bisphosphate aldolase class I
fructose bisphosphate aldolase class II
fructose bisphosphate aldolase monomer (polypeptide)
fructose bisphosphate aldolase monomer (polypeptide)
glycine hydroxymethyltransferase (serine aldolase)
L-allo-threonine aldolase
L-fuculose-phosphate aldolase
N-acetylneuraminate lyase (N-acetylneuraminic acid aldolase)
protein with similarity to HHED aldolases
Select this enzyme
Internet



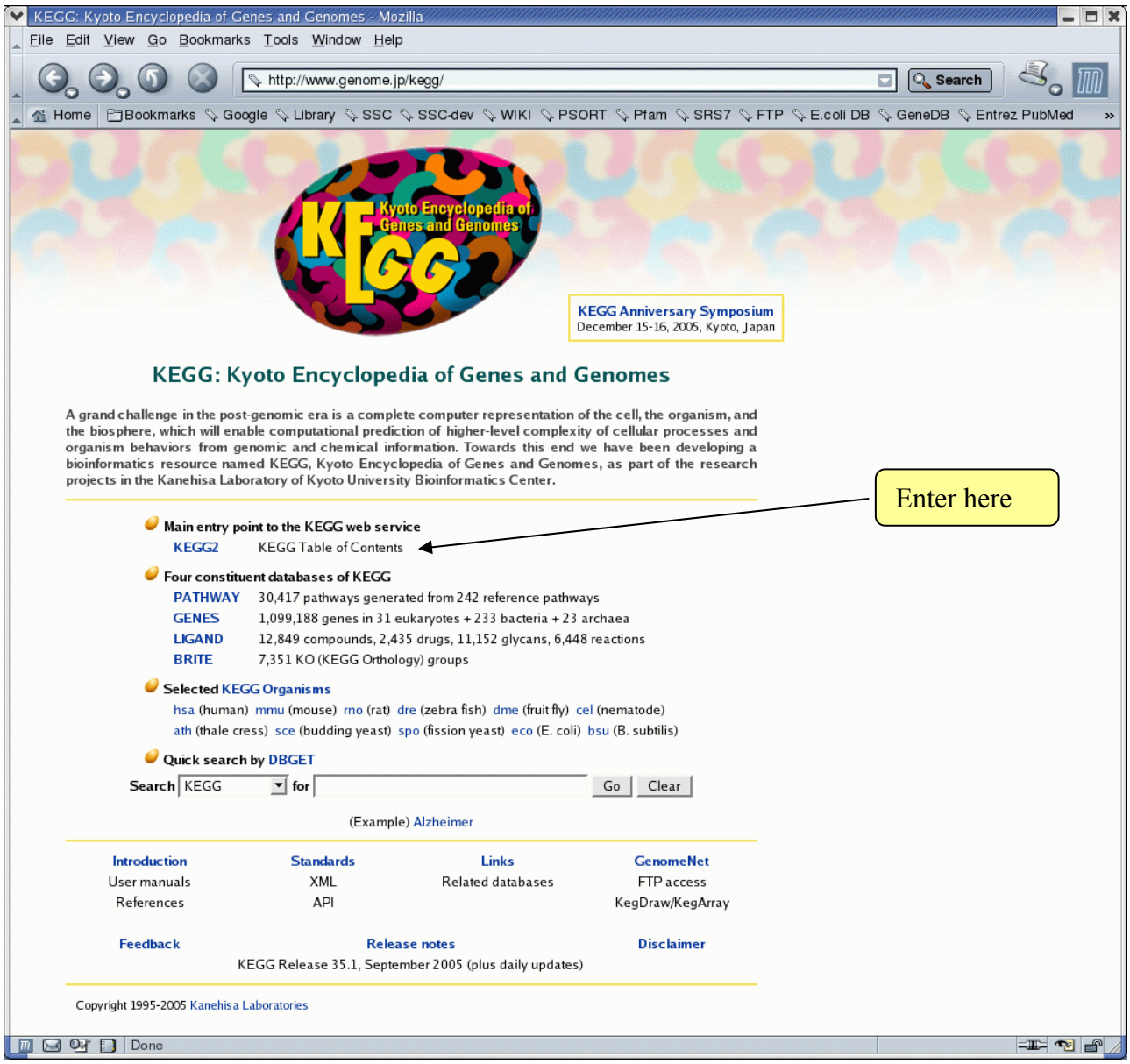
Exercise 3

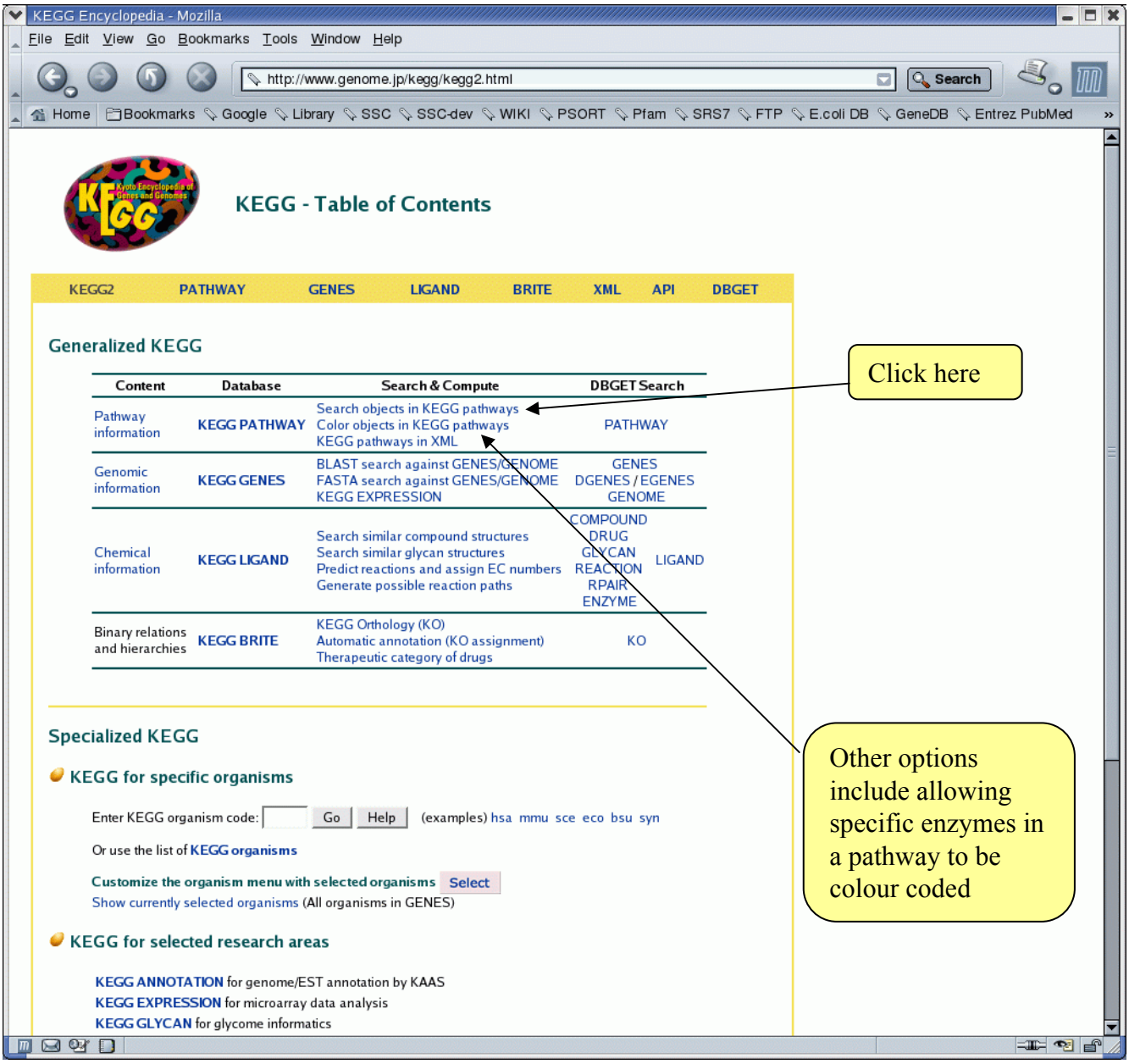
3. Here are six EC numbers for proteins in the malaria genome annotation that have been assigned based on protein similarities.
Are they involved in a common pathway?
If so, can you use KEGG to piece together the pathway and predict which gene is missing and therefore could remain unidentified in the Malaria genome.

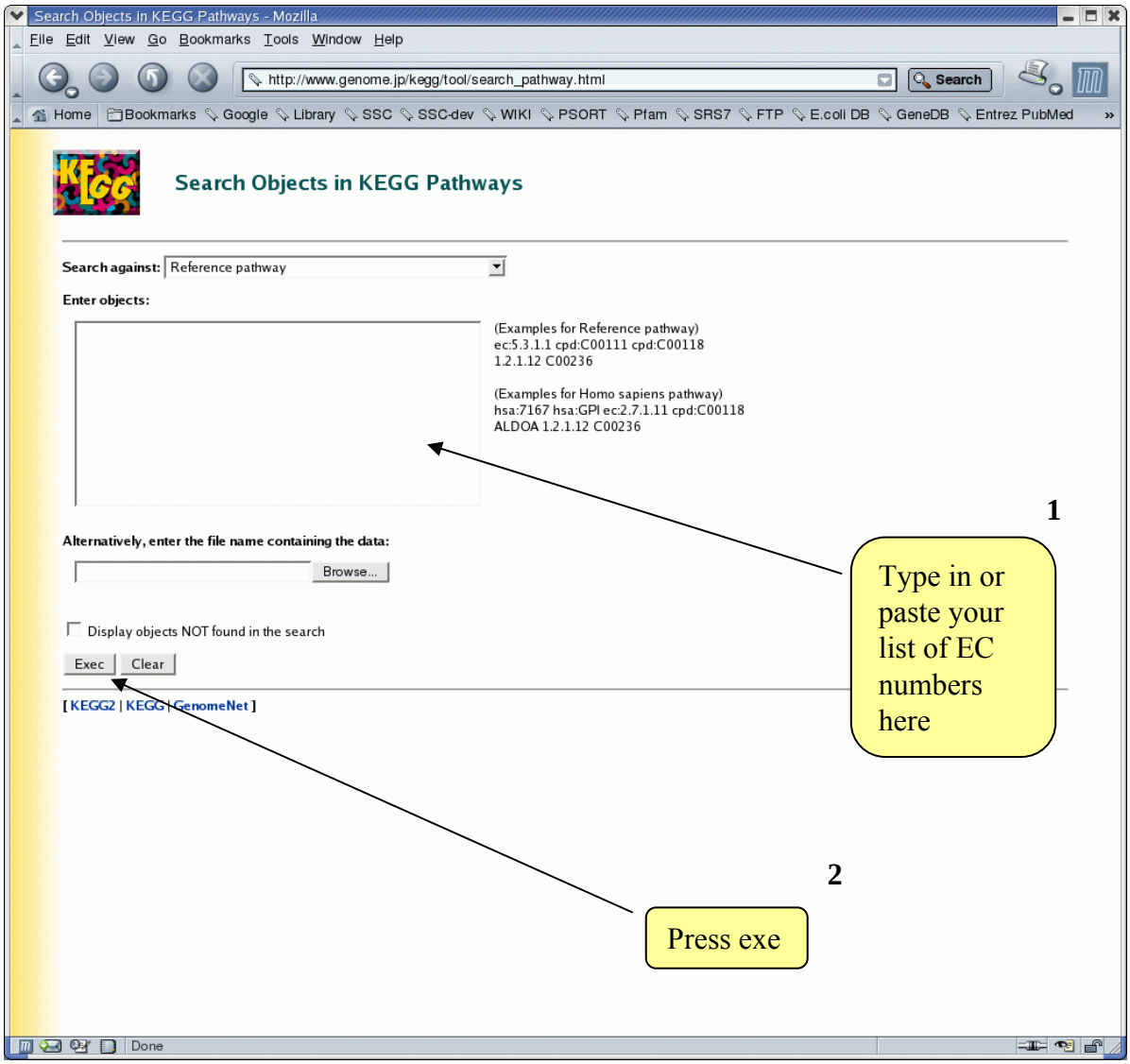
The first EC number is for a fructose bisphosphate aldolase and you already have it (see previous exercise).

The remaining known EC numbers are listed below:
2.7.1.1
5.4.2.8
2.7.7.13
4.2.1.47
2.7.1.90

Next, use KEGG to find possible pathways in which all these enzymes are found.
http:/www.genome.jp/kegg

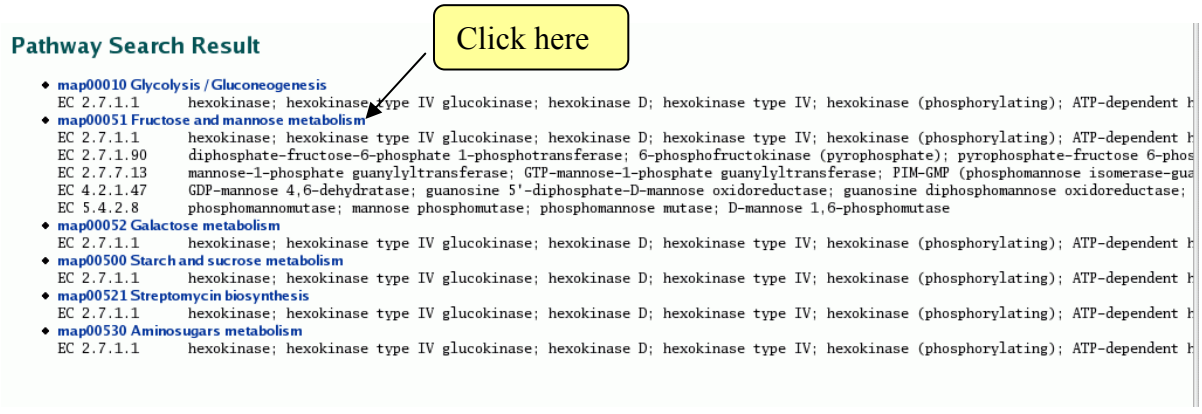


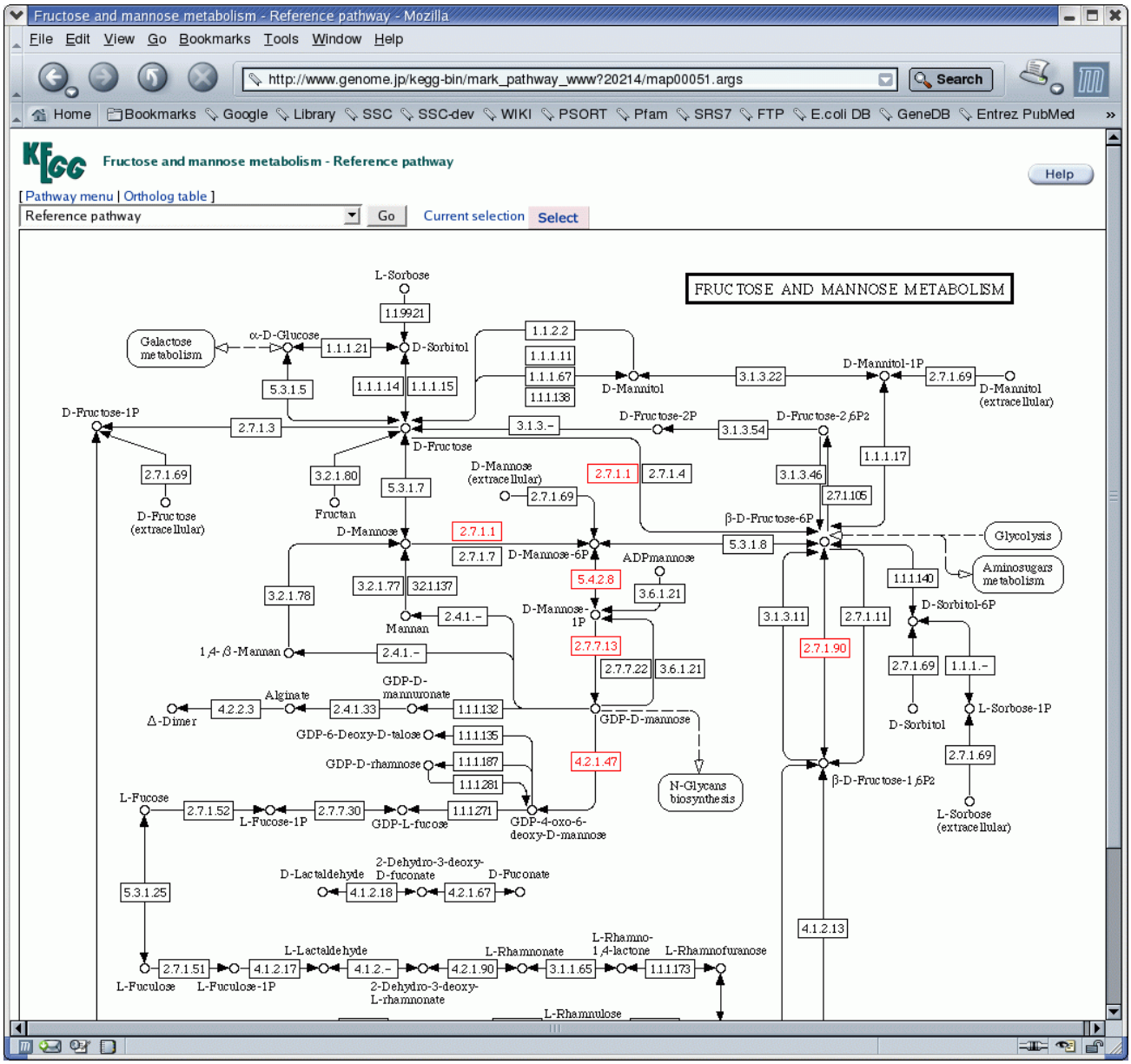

KEGG Encyclopedia - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://www.genome.jp/kegg/kegg2.html
Search
Home Bookmarks Google Library SSC SSC-dev WIKI PSORT Pfam SRS7 FTP E.coli DB GeneDB Entrez PubMed
KEGG - Table of Contents
KEGG2 PATHWAY GENES LIGAND BRITE XML API DBGET
Generalized KEGG
Content Database Search & Compute DBGET Search
Pathway information KEGG PATHWAY Search objects in KEGG pathways Color objects in KEGG pathways KEGG pathways in XML PATHWAY
Genomic information KEGG GENES BLAST search against GENES/GENOME FASTA search against GENES/GENOME KEGG EXPRESSION GENES DGENES / EGENES GENOME
Chemical information KEGG LIGAND Search similar compound structures Search similar glycan structures Predict reactions and assign EC numbers Generate possible reaction paths COMPOUND DRUG GLYCAN REACTION RPAIR ENZYME LIGAND
Binary relations and hierarchies KEGG BRITE KEGG Orthology (KO) Automatic annotation (KO assignment) Therapeutic category of drugs KO
Specialized KEGG
KEGG for specific organisms
Enter KEGG organism code: Go Help (examples) hsa mmu sce eco bsu syn
Or use the list of KEGG organisms
Customize the organism menu with selected organisms Select
Show currently selected organisms (All organisms in GENES)
KEGG for selected research areas
KEGG ANNOTATION for genome/EST annotation by KAAS
KEGG EXPRESSION for microarray data analysis
KEGG GLYCAN for glycome informatics
Click here
Other options include allowing specific enzymes in a pathway to be colour coded



In which pathway do all the enzymes from the list act?

Click on the map for that pathway. See over page.





The EC numbers of the enzymes you entered appear in Red.
Which enzyme might be missing? Click on it to view the KEGG entry.

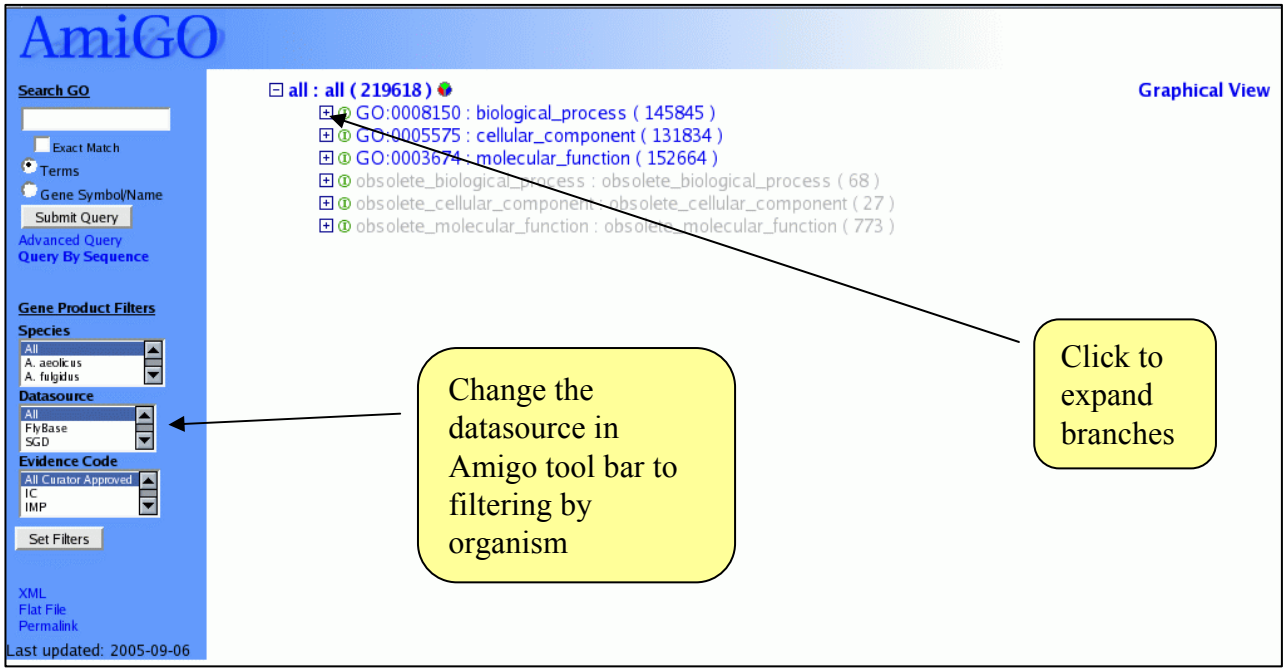
Hint – it joins the two pathways

Section 2

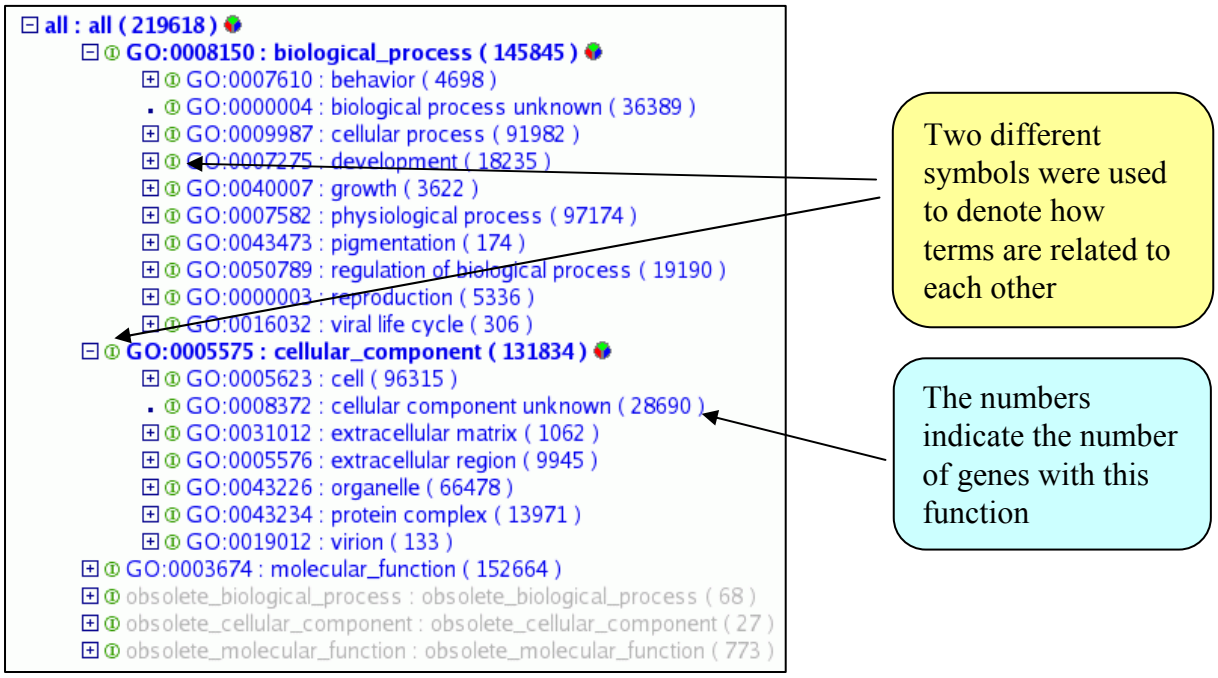
Gene Ontology (GO)

The official browser for GO annotations is "Amigo".
Web address: http://www.godatabase.org/

The structure of the ontology allows you to quickly find the description or term that you're looking for. The 'tree' of GO terms describes various facts of the proteins, function, cellular location etc. AmiGO holds all the gene predictions from multiple genomes within this tree. First, navigate through the "tree" by expanding and collapsing branches…

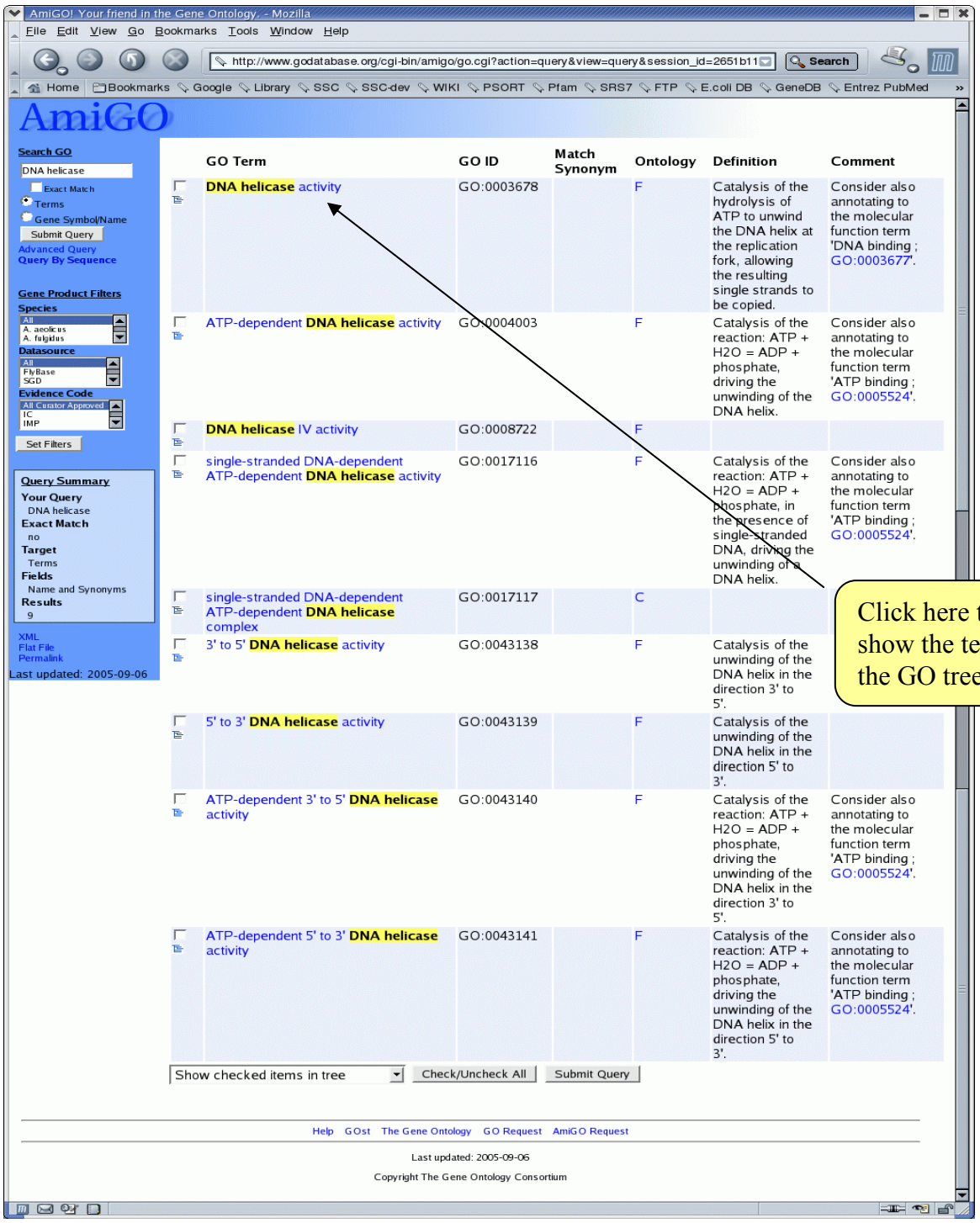


Here's an expanded view of biological process and cellular component.



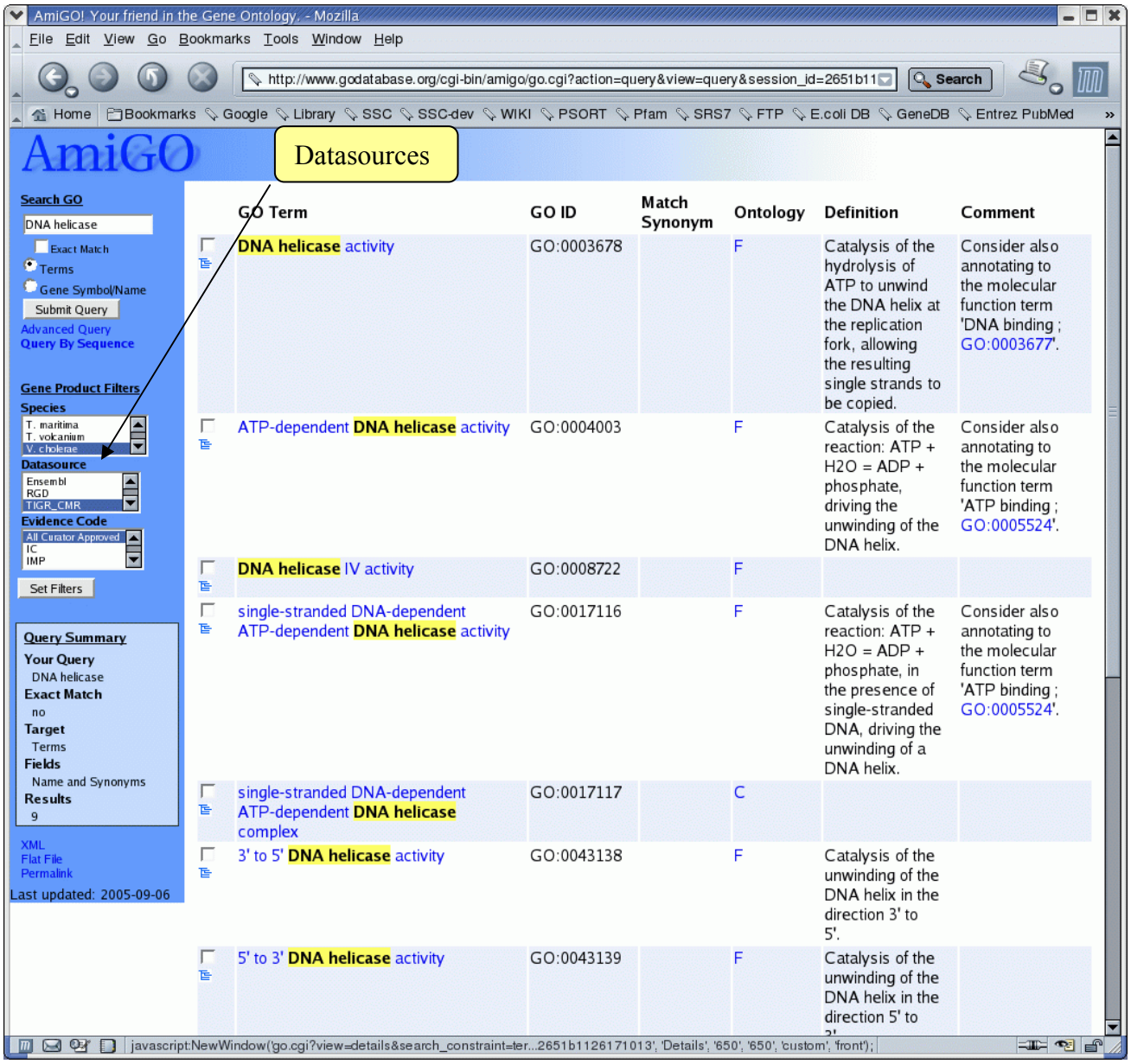
Exercise 1

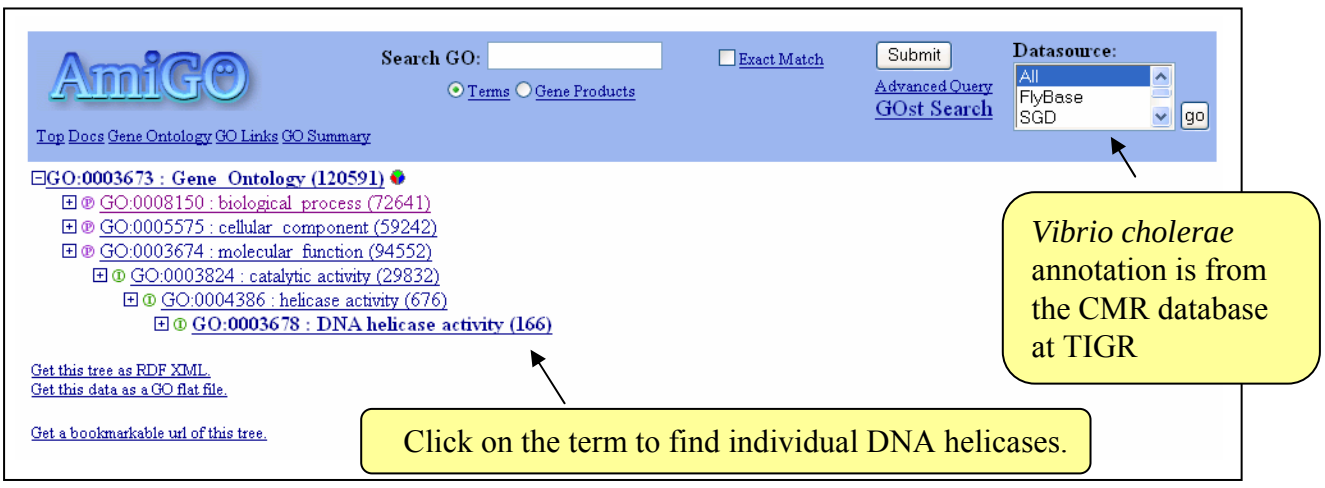
Next use the "search GO" box search for the term "DNA helicase“ and submit. several GO terms will appear.



Exercise II

By selecting an individual datasource, see how many *Vibrio cholerae* sequences have been annotated (see over page)





In the recent malaria genome paper fatty acid biosynthesis was highlighted as a possible target for chemotherapy.

Using Amigo:

Find the proteins involved in fatty acid biosynthesis in malaria.

Where are many of them they localised within the cell?

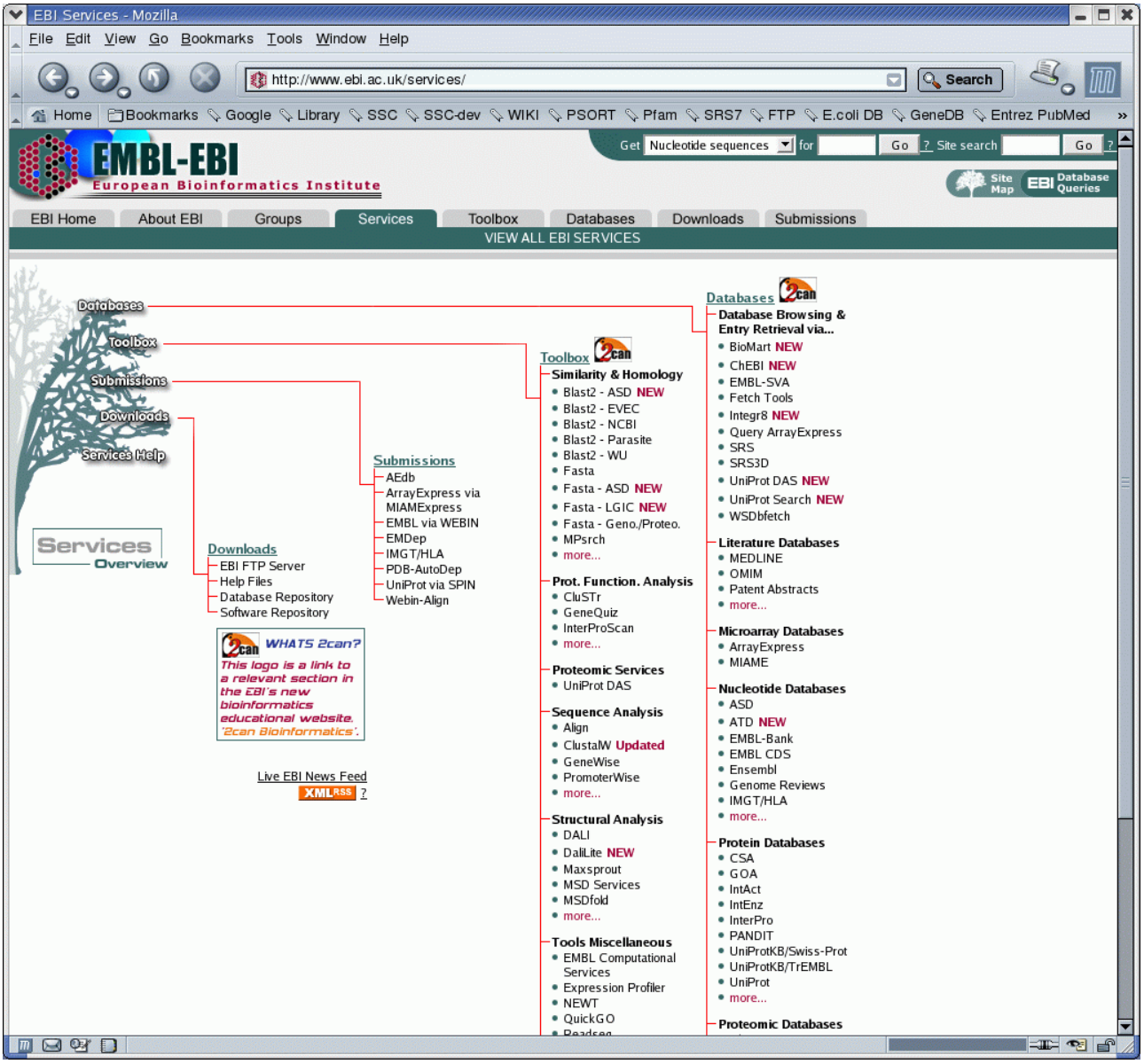
Are there any uncharacterised genes with the same subcellular localisation?

(NB. You can filter using the evidence code to show only those that have had their localisation experimentally confirmed). Ask if you are unclear.

Section 3

InterPro & UniProt

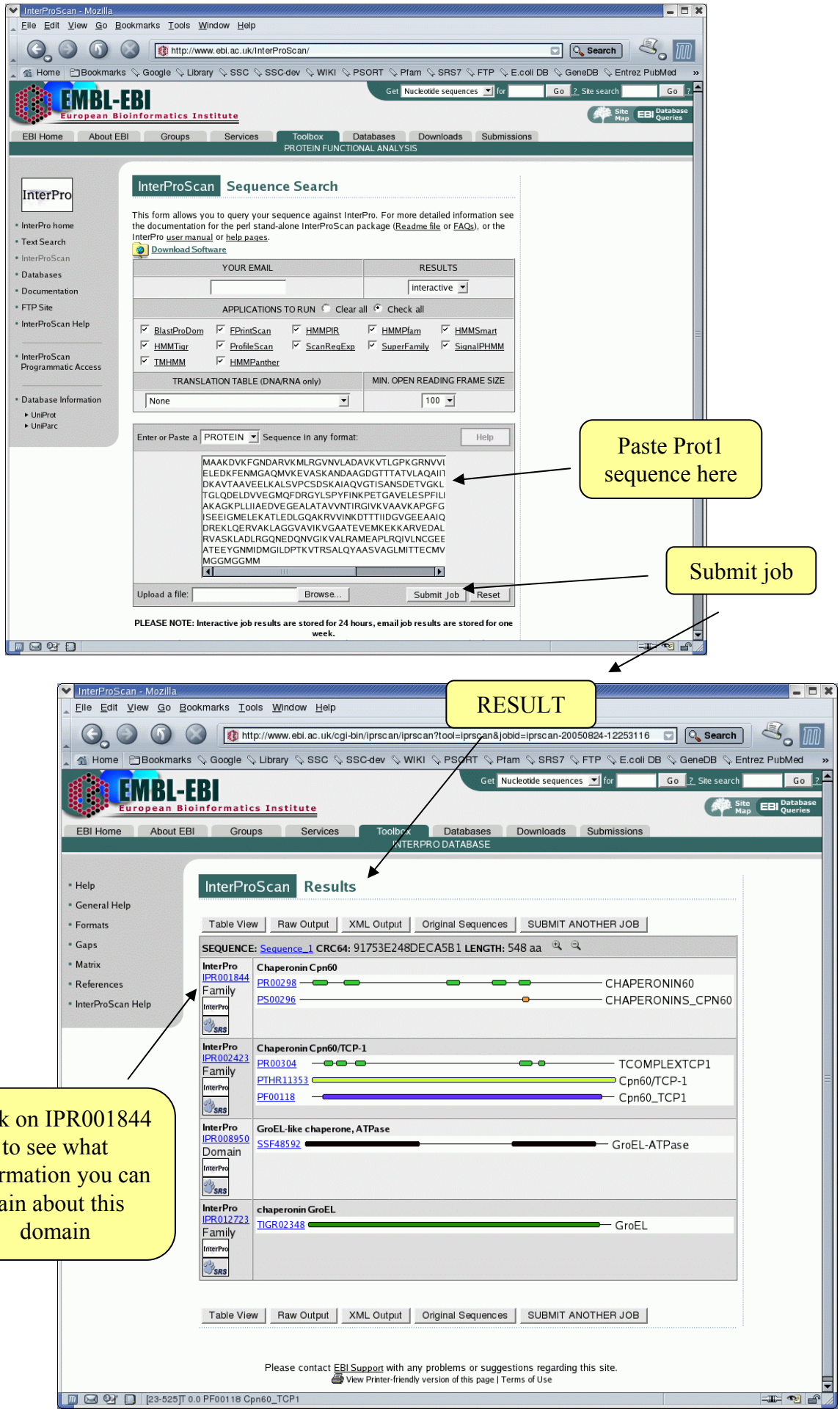
Web address: http://www.ebi.ac.uk/services



Exercise 1

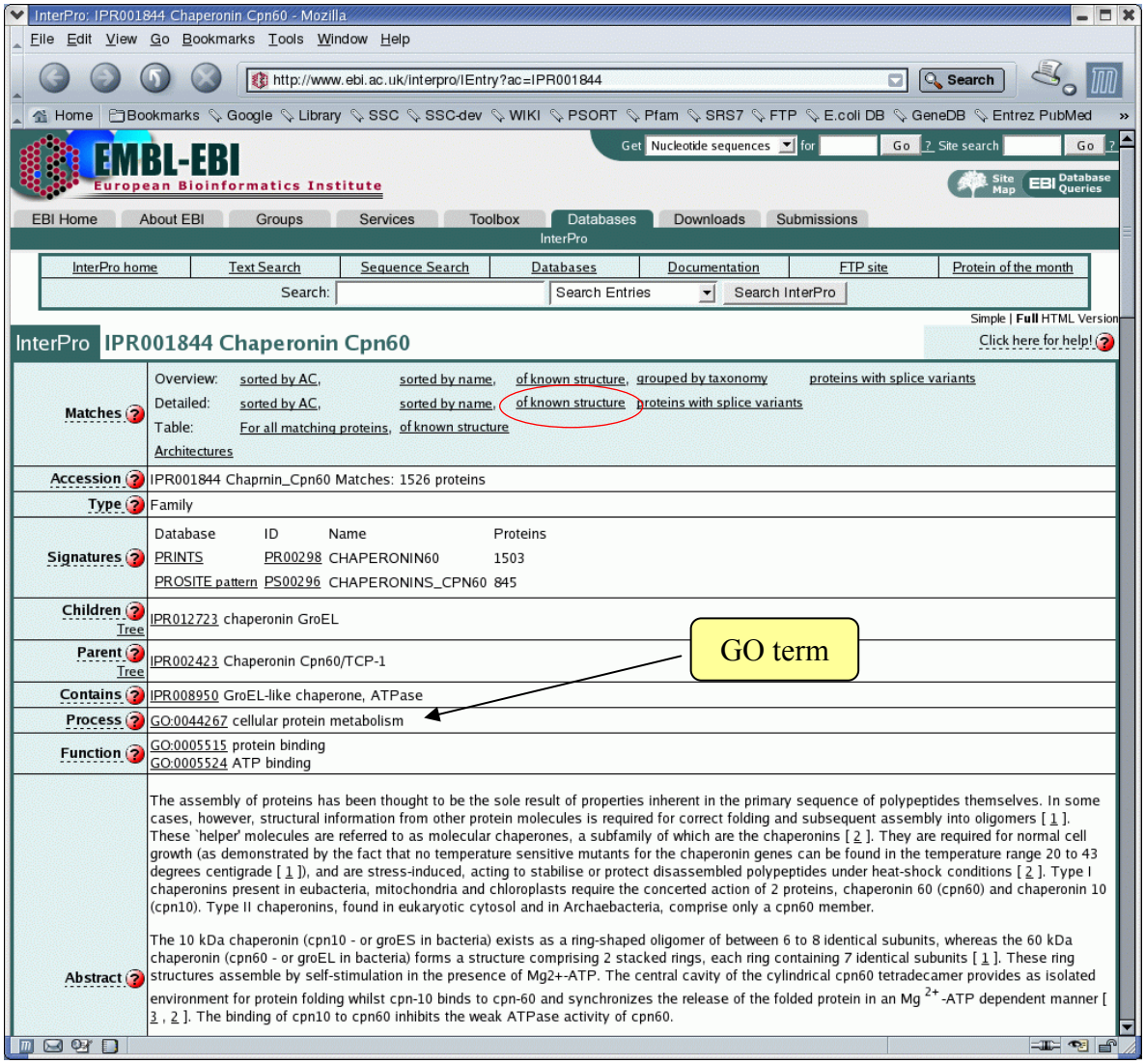
Use InteProScan to assign family membership and identify functional domains.

Copy and paste ‘Prot1’ amino acid sequence into the main box and submit job. You can find the sequence in the “sequences” file under the “Module_5” directory


InterProScan - Mozilla
EMBL-EBI
European Bioinformatics Institute
InterProScan Sequence Search
Paste Prot1 sequence here
Submit job
RESULT
InterProScan Results
Click on IPR001844 to see what information you can gain about this domain

Looking at the window from the page before:

•What domains/sites does this protein contain?
• Click on IPR001844 to see what information you can gain about this domain and look for GO terms that could be assigned to it.
• Scroll down the web page and look for the relationships that this entry might have with other InterPro entries.



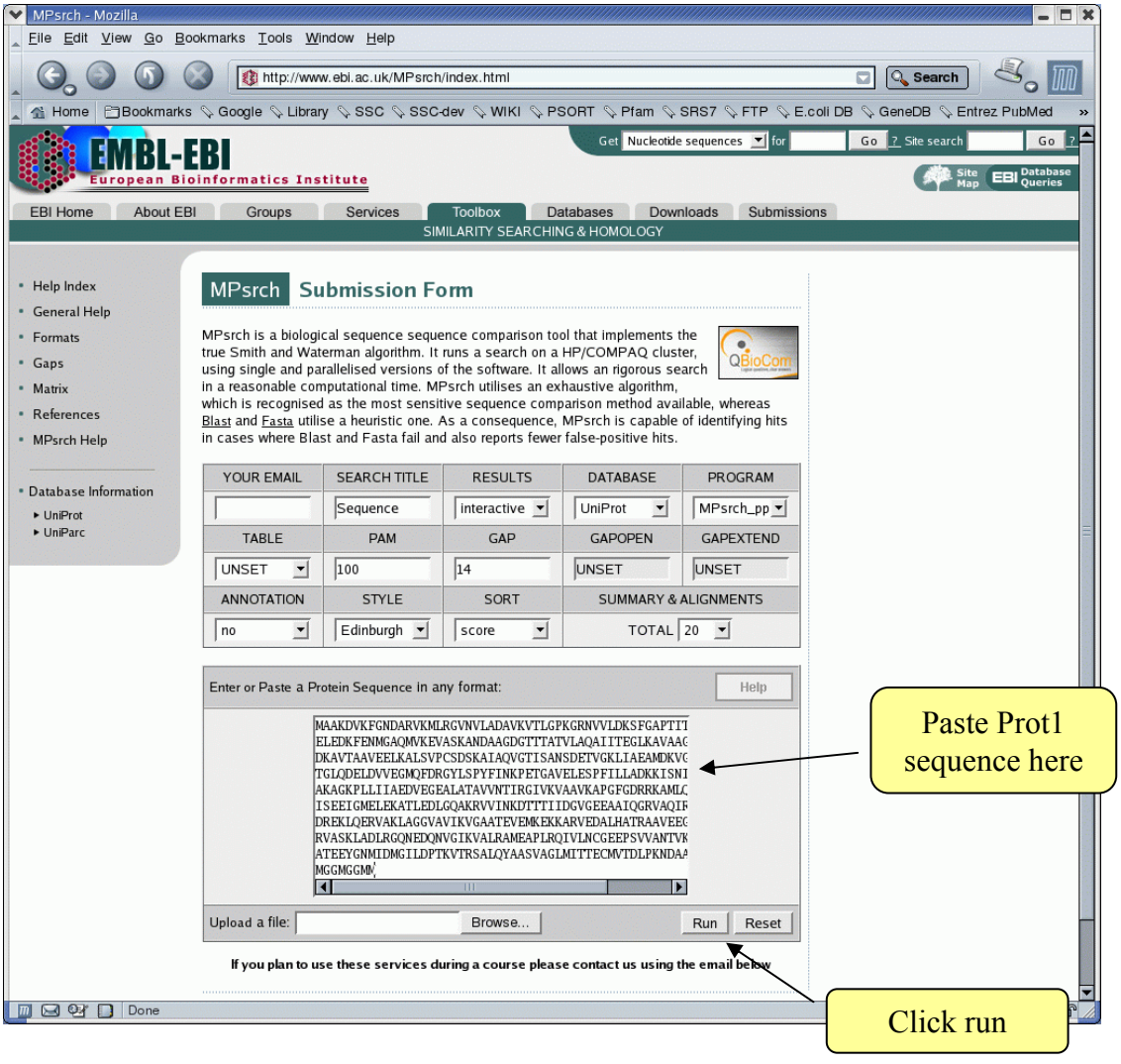
If you look at the top of the page, on the "*Detailed view*" line, follow the link to "*of known structure*" and look at the "*Structural features*" on the first protein (red circle). Note that it has a PDB structure (green stripped bar) for whole length of the protein and two ways of classifying the same domain: CATH (pink stripped bars) and SCOP (black stripped bars). Click on the links to see the differences between these two databases.

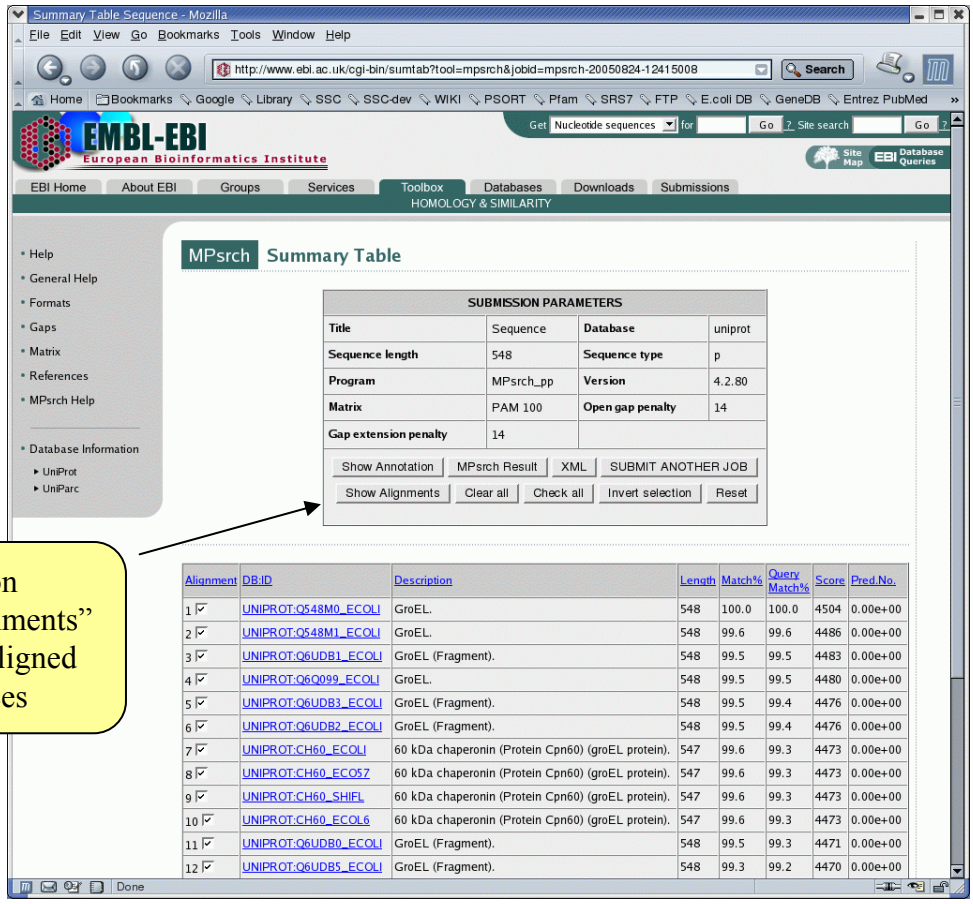
Click on the  symbol adjacent to the CATH domain to have a look at the domain using Astexviewer. Note that the selected CATH domain is highlighted in yellow.

Exercise 2

Alignments: MPsrch is capable of identifying hits in cases where Blast and Fasta fail.
Web page: http://www.ebi.ac.uk/MPsrch/index.html

Copy and paste Prot1 amino acid sequence into the main box and click run. You can find the sequence in the "sequences" file under the "Module_5" directory

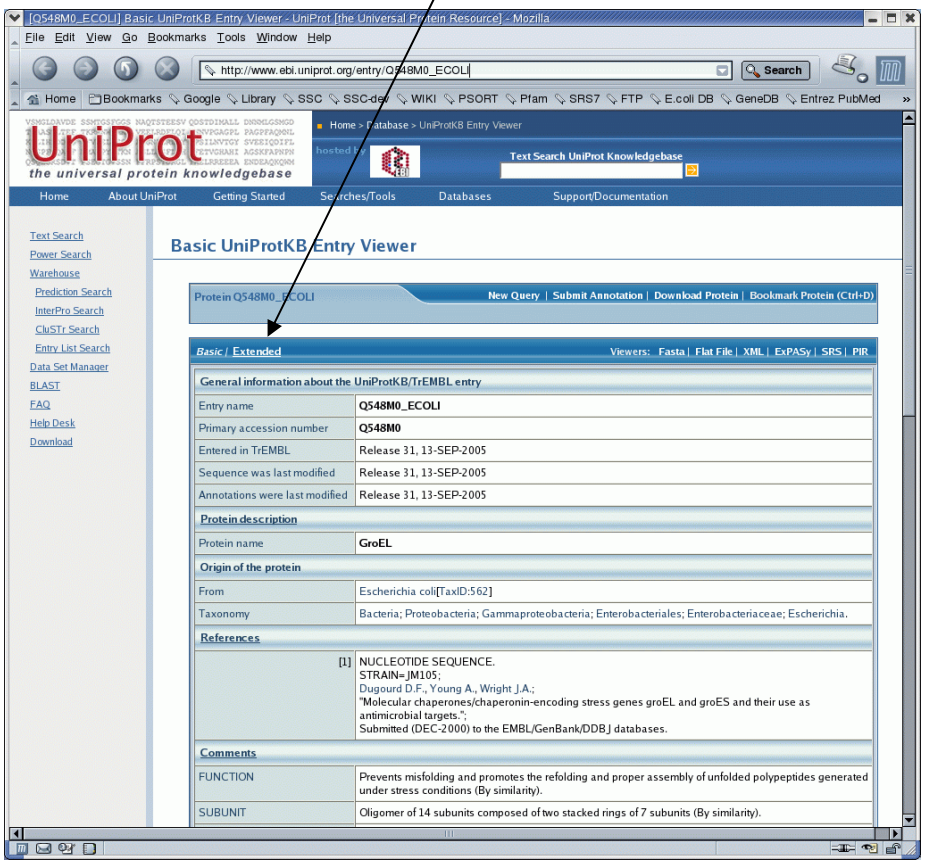




Click on "Show Alignments" to display aligned sequences

Alignment	DB:ID	Description	Length	Match%	Query Match%	Score	Pred.No.
1	UNIPROT:Q548M0_ECOLI	GroEL.	548	100.0	100.0	4504	0.00e+00
2	UNIPROT:Q548M1_ECOLI	GroEL.	548	99.6	99.6	4486	0.00e+00
3	UNIPROT:Q6UDB1_ECOLI	GroEL (Fragment).	548	99.5	99.5	4483	0.00e+00
4	UNIPROT:Q6Q099_ECOLI	GroEL.	548	99.5	99.5	4480	0.00e+00
5	UNIPROT:Q6UDB3_ECOLI	GroEL (Fragment).	548	99.5	99.4	4476	0.00e+00
6	UNIPROT:Q6UDB2_ECOLI	GroEL (Fragment).	548	99.5	99.4	4476	0.00e+00
7	UNIPROT:CH60_ECOLI	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
8	UNIPROT:CH60_ECO57	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
9	UNIPROT:CH60_SHIFL	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
10	UNIPROT:CH60_ECOL6	60 kDa chaperonin (Protein Cpn60) (groEL protein).	547	99.6	99.3	4473	0.00e+00
11	UNIPROT:Q6UDB0_ECOLI	GroEL (Fragment).	548	99.5	99.3	4471	0.00e+00
12	UNIPROT:Q6UDB5_ECOLI	GroEL (Fragment).	548	99.3	99.2	4470	0.00e+00

You can click on the UniProt links for each entry and browse through the different links you find in both the 'Basic' and the 'Extended' web pages.



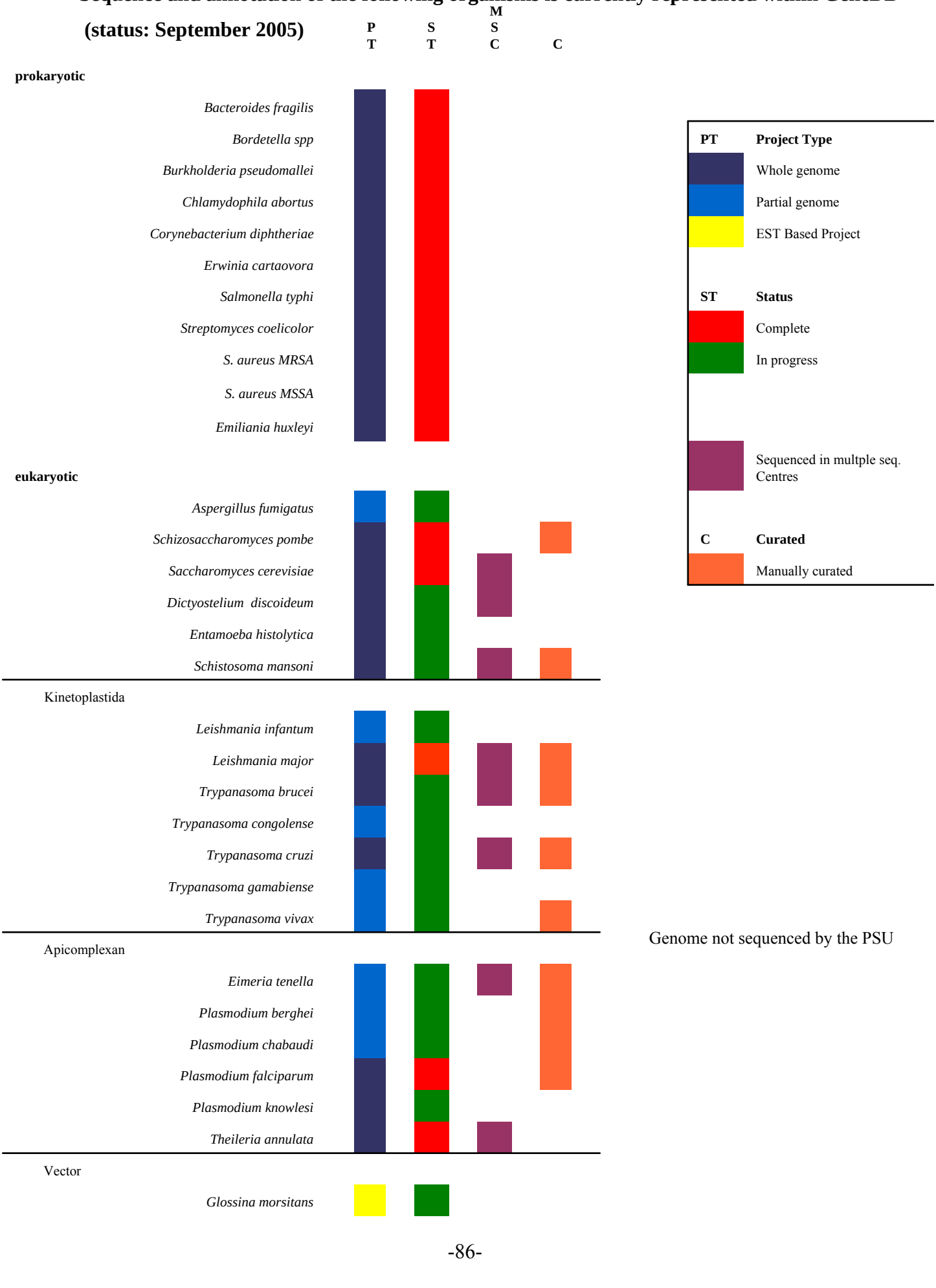
Module 6
Data mining using GeneDB

Introduction

This module will demonstrate GeneDB (http://www.genedb.org), a genome database housing sequence and annotation of prokaryotic and eukaryotic organisms. The resource provides a portal through which data generated by the Pathogen Sequencing Unit and other collaborating sequencing centres can be made publicly available. It combines data from finished and ongoing genome and expressed sequence tag (EST) projects with curated annotation, that can be searched, sorted and downloaded, using a single web based resource. The current release stores 33 datasets (see Table 1) of which 12 are curated and maintained by biologists, who review and incorporate information from the scientific literature, public databases and the respective research communities.

Sequence and annotation of the following organisms is currently represented within GeneDB

(status: September 2005)



Aims

The aim of this module is for you to familiarise yourself with GeneDB and the various ways of accessing, querying, browsing and retrieving data. You'll use GeneDB as a research tool to retrieve candidate genes which you could follow up with further experimental validation. In the process you will also hopefully see how GeneDB integrates diverse biological datasets, organises, indexes and extensively cross-references these.

In addition, the exercises are designed to make more general points which need to be taken into consideration when approaching and evaluating database searches, not just GeneDB. These are:

1. How complete/incomplete is the dataset you are searching? In the case of organisms with two sets of each chromosome (i.e. diploid organisms), does the dataset represent the haploid or diploid genome content?

2. How was the dataset generated?
    a. Is it an EST project? What estimated coverage does the dataset represent (i.e. is it a partial 3-5x coverage or a 8-10x coverage project?)
    b. Has the sequence been "manually" finished (i.e. sequencing gaps closed and base checked)?
    c. How were the gene predictions carried out (automated vs. manual)?
    d. How were the gene prediction annotated (automated vs. manual)?

3. Depending on the gene prediction and associated annotation method, you may need to approach querying from several angles, not just 1 methodology (e.g. combine keyword searching with similarity searching).

4. When designing your searches:
    a. identify keywords that describe your topic.
    b. identify any synonyms for your keywords.
    c. be aware of spelling variations and/or plurals.
    d. decide the scope of your search.
    e. be aware that using the same search method in different databases may affect your results.
    f. try different search methods to identify candidate genes.
    g. be aware of the use of wildcards.

Exercise 1:
Data mining the *T. brucei* genome for the Arp2/3 complex.
Exercise 2:
Using the Artemis applet to retrieve sequence and annotated features.
Exercise 3:
Demonstration of the Boolean querying tool.
Exercise 4:
Data mining of *Plasmodium* genomes for monosaccharide transporters.
Exercise 5:
Data mining three *Bordetella* genomes for autotransporter genes

Exercise 1 Data mining of the *T. brucei* genome for the Arp2/3 complex.

Can you identify the components of the Arp2/3 complex in the kinetoplastid organism *Trypanosoma brucei*, causative agent of sleeping sickness in sub-Saharan Africa?

The Arp2/3 complex is involved in actin assembly and function in the eukaryotic cytoskeleton. So far this complex has not been investigated in kinetoplastids, but has been well characterised in other organisms, such as the fission yeast *Schizosaccharomyces pombe* . Unlike the *S. pombe* genome, which is complete and contains extensive curated annotations, the genomes of the trypanosomatids are in various stages of completion and annotation. Using a variety of tools/methods, some of which you will already have covered in earlier modules, identify putative members of this complex and complete the table on page at the end of this exercise.

Start by identifying how many components have been annotated to this complex in *S. pombe*, which you will be using as a thoroughly annotated reference genome (**exercise 1.1**).

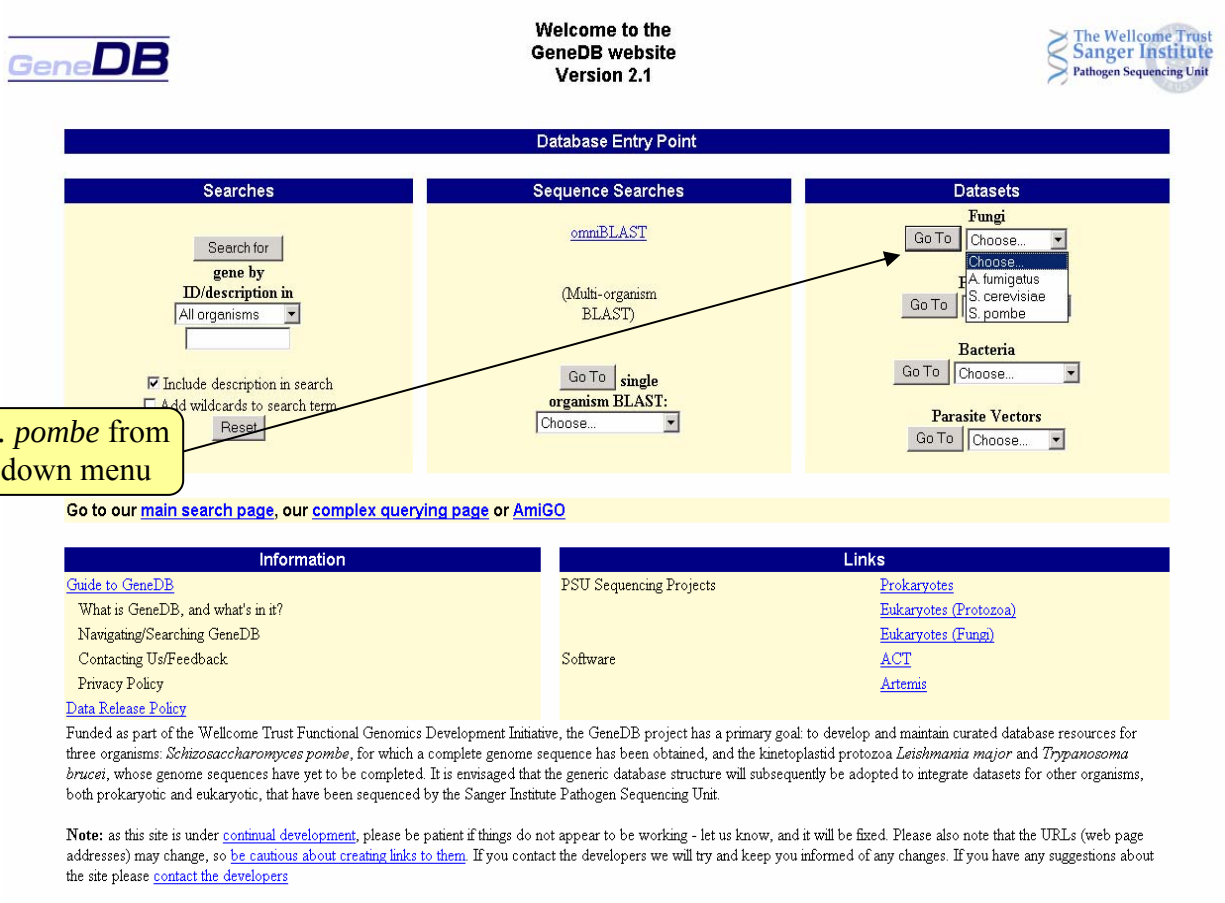
Exercise 1.1

1

Go to the GeneDB homepage (http://www.genedb.org)

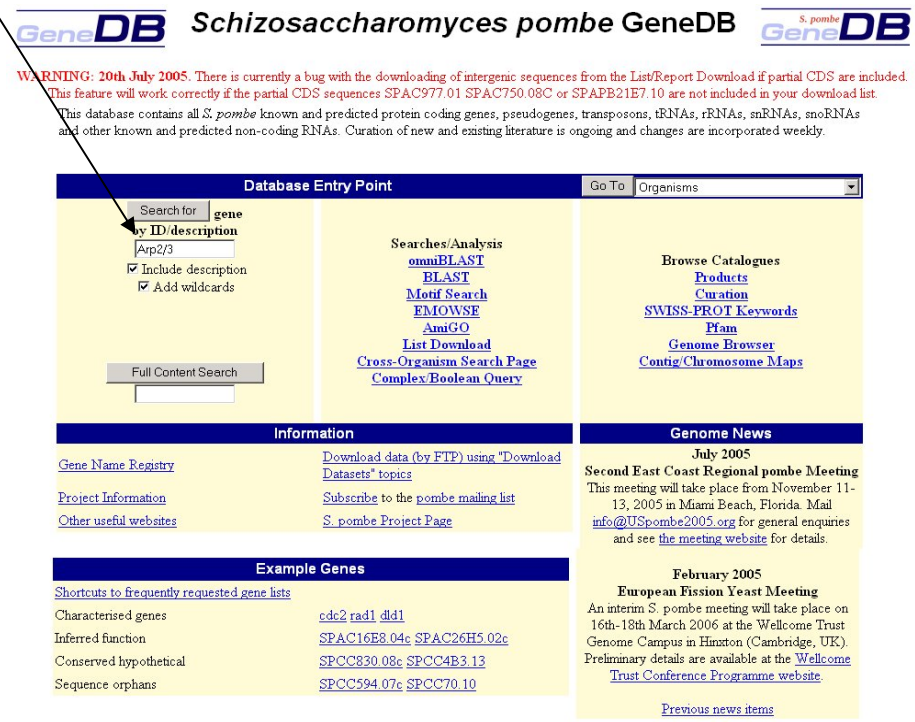
2

Select *S. pombe* from the pull down menu



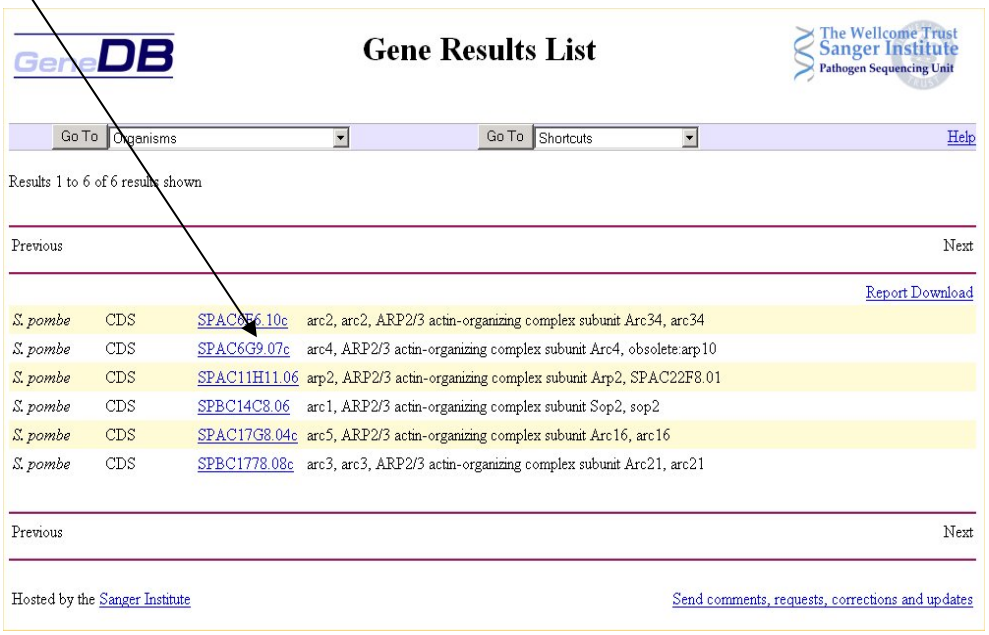
3

Type **Arp2/3** into the search box, ensuring that the '**Add wildcards to search term**' box is also selected. This will search all the gene names and product/description lines assigned to features within the *S. pombe* dataset. Not selecting the wildcard box would only return exact matches.



4

The results list below will be returned. Clicking on the hyperlinked gene names will take you to the feature page. Click on SPAC6G9.07c to see the information contained on a feature page.



5

Click on the links on the feature page to see how the data are cross-linked and referenced.

GeneDB

CDS: arc4

S. pombe GeneDB

WARNING: **20th July 2005.** There is currently a bug with the downloading of intergenic sequences from the List/Report Download if partial CDS are included. This feature will work correctly if the partial CDS sequences SPAC977.01 SPAC750.08C or SPAPB21E7.10 are not included in your download list.

Search for | Go To Organisms | Go To Shortcuts | Help | Contact curator

Biological process, cellular component and molecular function annotation is being removed from the description part gene pages and replaced by manual GO curation. To obtain the full complement of gene products annotated to specific terms please use the links to Amigo in the left hand column of the "Gene Ontology" section of the page.

General Information Add to Basket | View Basket

Name	arc4
Systematic Name	SPAC6G9.07c
Obsolete Synonyms	arp10
Status	role inferred from homology
Product	ARP2/3 actin-organizing complex subunit Arc4
Type	CDS
Sequence	DNA and Protein

Location

Chromosome	1
Contig	c977
Contig Location	complement(3228809..3229681) (Unspliced length: 873 bp)
Exons	complement(join(3228809..3228983 , 3229167..3229290 , 3229421..3229623 , 3229679..3229681)) (Spliced length: 507 bp)

Graphical Display (in Artemis) | Genome Browser

Context Map:

SPAC6G9.03c SPAC6G9.04 SPAC6G9.05 pcp1 >arc4< ubp6 rpl24 sen1 syb1 cfr1

Curation

ARP2/3 actin-organizing complex subunit Arc4
similar to S. cerevisiae YKL013C

Predicted Peptide Properties

Mass	19.6 kDa	Amino acids	168
Isoelectric point	pH 5.2	Charge	-3.0

Signal Peptide	Not found
Transmembrane Domains	0 found
GPI Anchor	Not found

Gene Ontology Annotation

Term (browse Amigo)	Qualifier	Evidence	Other genes annotated to this term
Biological Process			
actin cortical patch assembly		ISS (GOC:unpublished) with SPTR:P33204	15 others
regulation of actin filament polymerization		ISS (GOC:unpublished) with SPTR:P33204	4 others
Cellular Component			
actin cortical patch		ISS (GOC:unpublished) with SPTR:P33204	41 others
Arp2/3 protein complex		ISS (GOC:unpublished) with SPTR:P33204	6 others

Published Expression Profiles

Gene Expression Viewer Cell Cycle Meiosis Environmental Stress

Phenotype

FYSSION (*S pombe* strain database)

Literature

Search for arc4 in PubMed

Domain Information

View Pfam domain structure for this gene product
View SCOP superfamily

DB	Accs	Description
Pfam	PF05856	ARP2/3 complex 20 kDa subunit (ARPC4)
InterPro	IPR008384	ARP23 complex 20 kDa subunit

Database Cross-References

DB	Accs	Description
UniProt	Q92352	Probable ARP2/3 complex 20 kDa subunit (p20-ARC).
EMBL	Z81317	S.pombe chromosome I cosmid c6G9.
EMBL	AB010050	Schizosaccharomyces pombe mRNA for 20 kd actin related protein complex, partial cds.
GermOnline	SPAC6G9.07c	GermOnline
PIR	T39069	PIR
PIR	T43309	PIR
PombePD	SPAC6G9.07c	Proteome, Inc

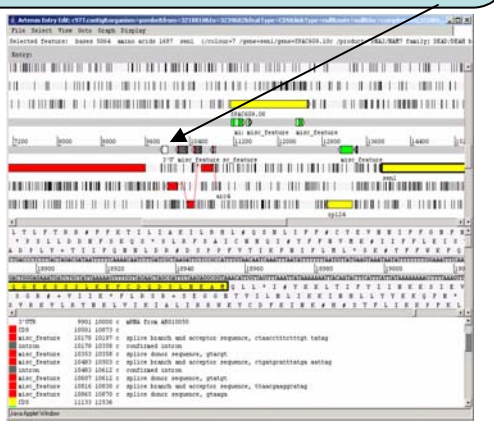
SWISS-PROT Annotation For This Protein

Similarity	Belongs to the ARPC4 family.
Function	Part of a complex implicated in the control of actin polymerization in cells (By similarity).
Sub Unit	Belongs to a complex composed of ARP2, ARP3, P41-ARC, P34-ARC, P21-ARC, P20-ARC and P16-ARC (By similarity).
Keywords	Complete proteome (4969 others), Cytoskeleton (25 others)

Navigation bar pull down menues: You can navigate between different organism datasets and search tools using pull down menus

Gene name and product information: The product lines are standardized and indexed so that features sharing the same product lines can be retrieved. Access to the nucleotide and amino acid sequences of the feature are also provided.

Basic location information and context map: Clicking on the 'Graphical display in Artemis' open up an Artemis applet – which will be discussed further in exercise 2. Via the applet, the feature can be viewed in the context of the sequence and additional annotation, such as UTRs.

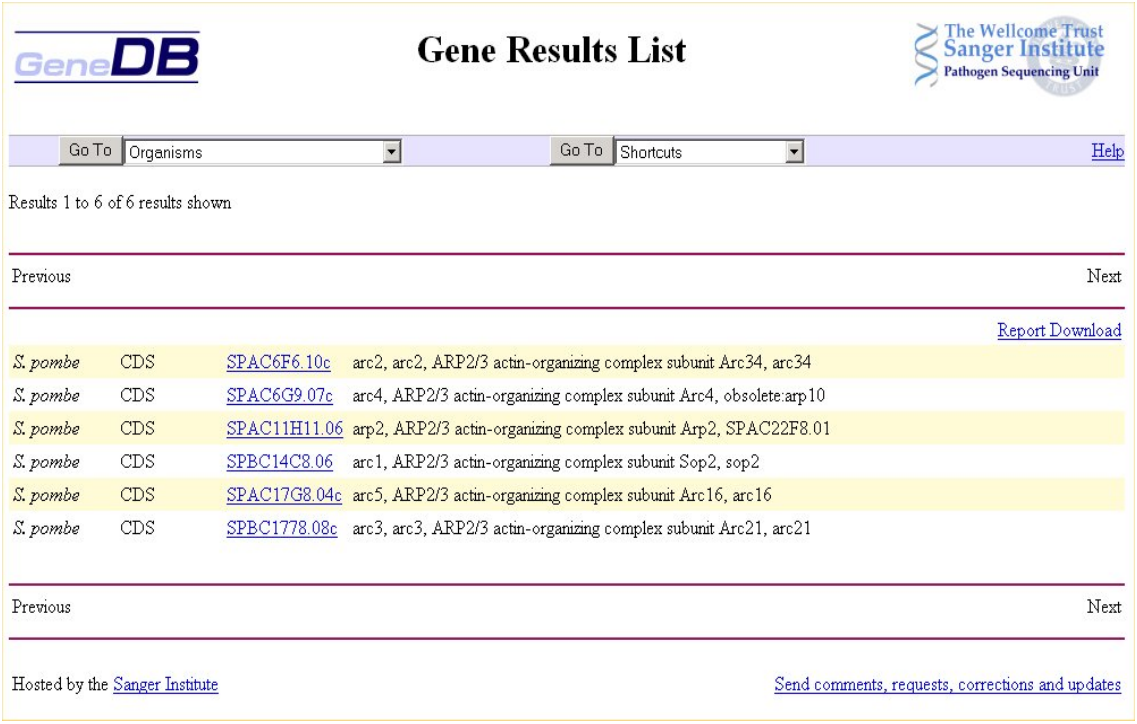


GO associations: Links will take you to the descriptions of the terms as well as other proteins annotated to the same ontology node.

Database cross-references to literature, phenotype, protein motif/domain as well as sequence databases.

6

Go back to your initial results page and see how many proteins have been assigned to the Arp2/3 complex in *S. pombe*. Are any of the components characterised by Pfam domains? Fill in the table on page at the end of the exercise.



GeneDB

Gene Results List

The Wellcome Trust Sanger Institute Pathogen Sequencing Unit

Go To Organisms | Go To Shortcuts | Help

Results 1 to 6 of 6 results shown

Previous Next

Report Download

S. pombe	CDS	SPAC6F6.10c	arc2, arc2, ARP2/3 actin-organizing complex subunit Arc34, arc34
S. pombe	CDS	SPAC6G9.07c	arc4, ARP2/3 actin-organizing complex subunit Arc4, obsolete:arp10
S. pombe	CDS	SPAC11H11.06	arp2, ARP2/3 actin-organizing complex subunit Arp2, SPAC22F8.01
S. pombe	CDS	SPBC14C8.06	arc1, ARP2/3 actin-organizing complex subunit Sop2, sop2
S. pombe	CDS	SPAC17G8.04c	arc5, ARP2/3 actin-organizing complex subunit Arc16, arc16
S. pombe	CDS	SPBC1778.08c	arc3, arc3, ARP2/3 actin-organizing complex subunit Arc21, arc21

Previous Next

Hosted by the Sanger Institute Send comments, requests, corrections and updates

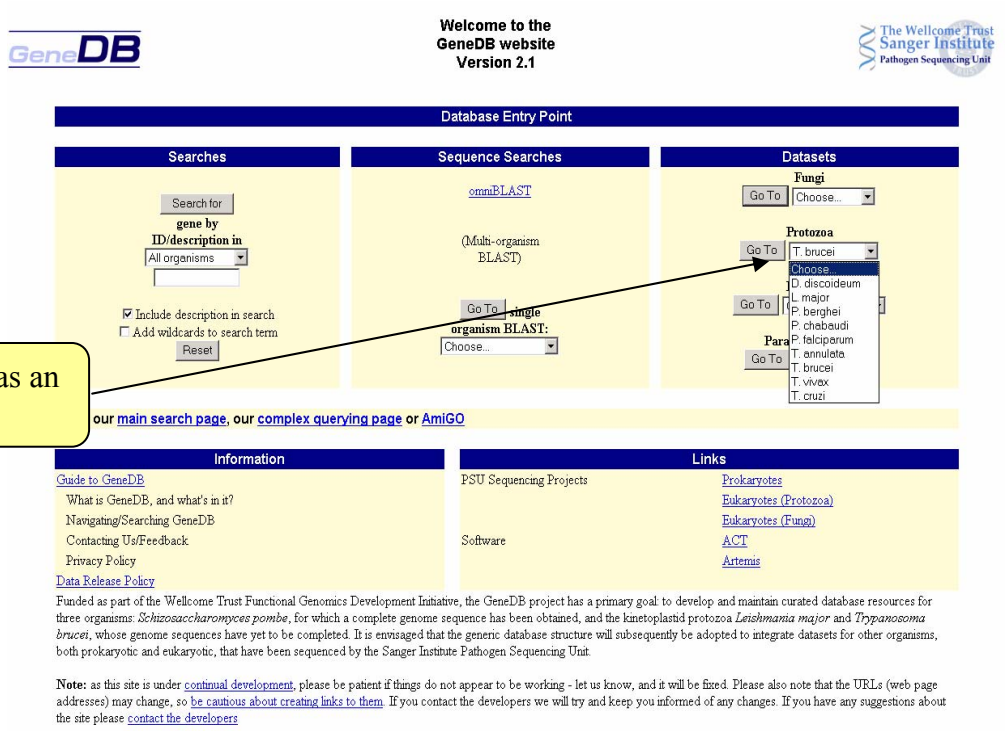
The next aim is to identify putative Arp2/3 complex components in the genomes of trypanosomatids. We're going to start off with the *T. brucei* genome. You may take a number of different approaches:

- using keywords to search the available *T. brucei* annotation (**exercise 1.2**).
- browsing predicted Pfam domain catalogue (**exercise 1.3**).
- using GO annotations and the gene ontology browser (**exercise 1.4**).
- using BLAST to identify sequences with similarity to known Arp2/3 components (**exercise 1.5)**.

Exercise 1.2 The use of keywords to search the available *T. brucei* annotation

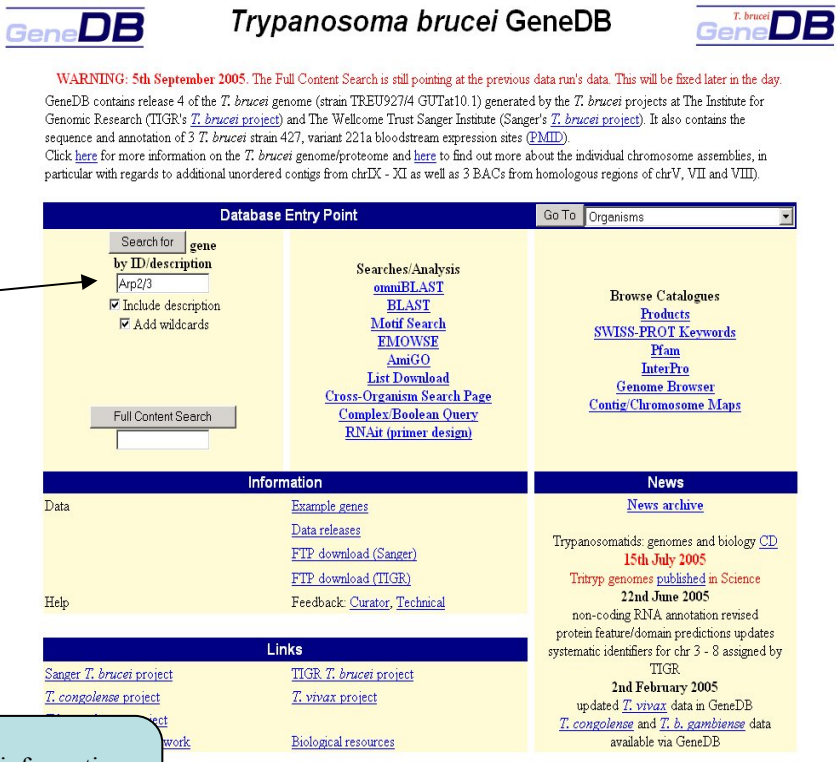
1

Go to the GeneDB homepage (http://www.genedb.org)



2

Select *T. brucei* as an organism



3

Type **Arp2/3** into the search box

The homepages also provide up-to-date information about sequencing progress, data updates, nomenclature and other community resources.

4

The results list below will be returned. Which components of the Arp2/3 complex have already been annotated in the *T. brucei* genome? Have a look by clicking on the hyperlinked gene names. Would you agree with the assignments?

Previous Next

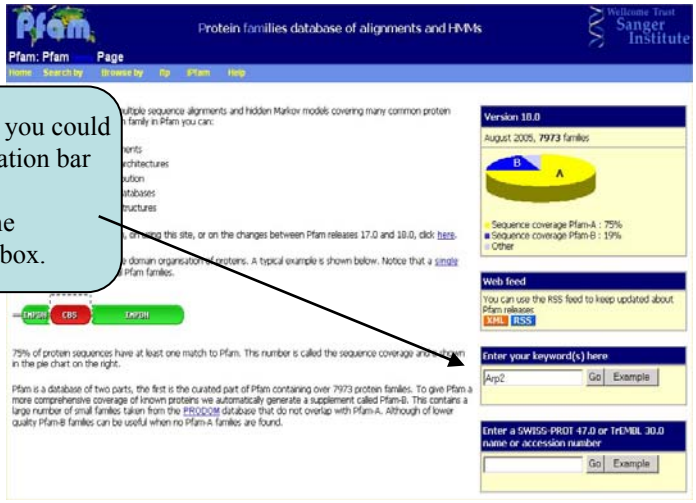
Report Download

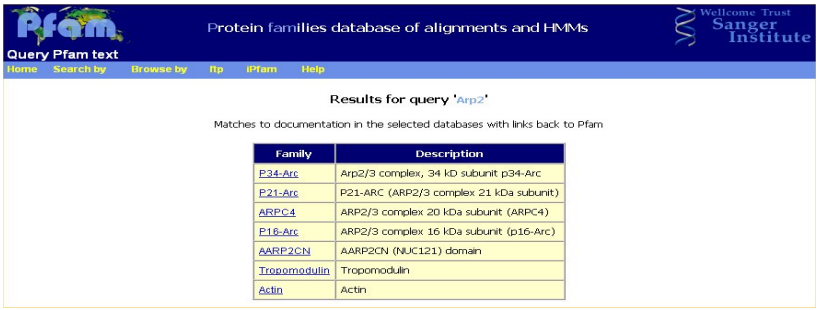
T. brucei	CDS	Tb10.70.2680	ARP2/3 complex subunit, putative, ARP2/3 complex subunit, putative, TRYP_x-70a06.p2kb545_360
T. brucei	CDS	Tb927.8.4410	ARP2/3 complex subunit, putative, Tb08.29H22.800
T. brucei	CDS	Tb10.389.0270	actin related protein 2/3 complex, putative, ARP2/3 complex subunit, putative
T. brucei	CDS	Tb10.406.0320	ARP2/3 complex 16kDa subunit, putative, ARP2/3 complex subunit, putative
T. brucei	CDS	Tb927.2.2900	ARP2/3 complex subunit, putative, 10C8.250

Exercise 1.3 Browsing the Pfam domain catalogue

As you will have seen from the *S. pombe* example earlier on, some of the subunits are characterised by Pfam domains and you should have made a note of these in the table. Therefore, one way of identifying the putative components would be by browsing the Pfam catalogue.

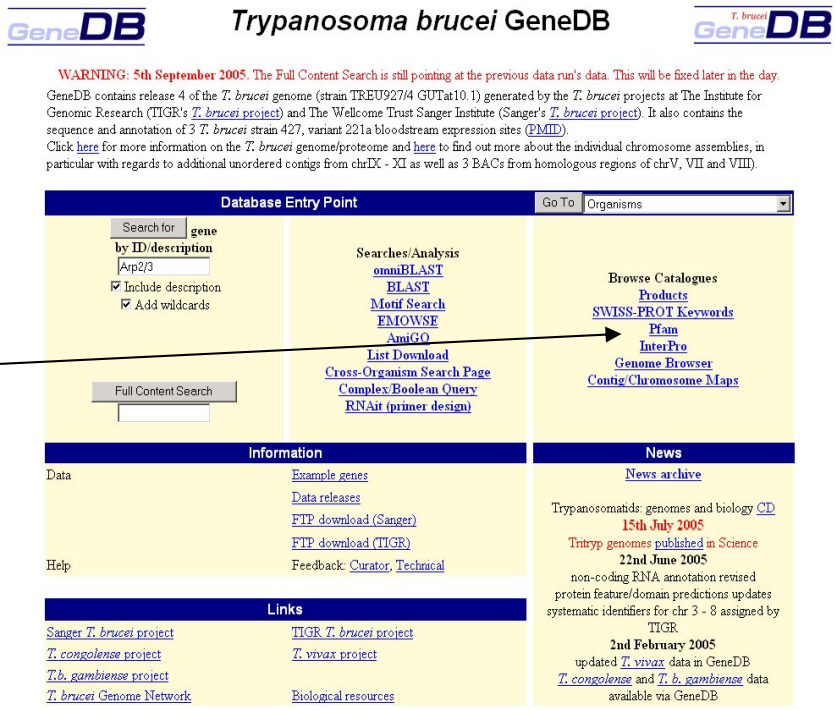
If you haven't made a note of the Pfam domains, then you could either go back to the *S. pombe* dataset using the navigation bar alternatively, use the Pfam site at http://www.sanger.ac.uk/Software/Pfam/ to retrieve the domain information by typing **Arp2/3** into the search box.





1

Select the Pfam link on the *T. brucei* homepage

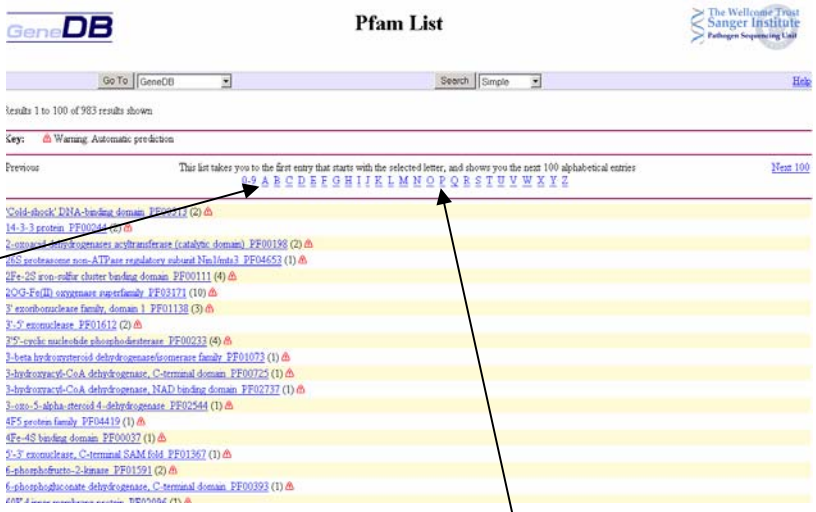


2

The Pfam domain descriptions will be listed alphabetically

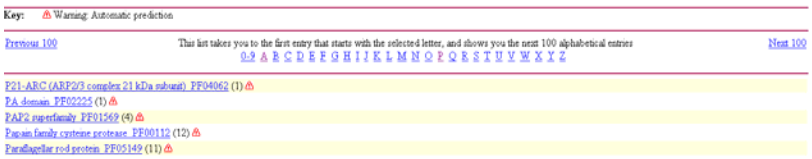
3

Click on the letter '**A**' and scroll down the list



4

Click on the letter '**P**'



How many of the putative Arp2/3 cpmplex components can be identified using the Pfam catalogue?

Exercise 1.4 Using GO annotations

By now, you should have identified putative homologues of 5 out of the 7 Arp2/3 complex components. You should be missing the 2 actin related proteins Arp2 and Arp3. Again, there are a number of ways to retrieve possible candidates:

• You could browse the Pfam catalogue for **Actin (PF00022)** – this will give you a short list of 9 candidates as a number of actin related proteins as well as actin itself which share this domain.

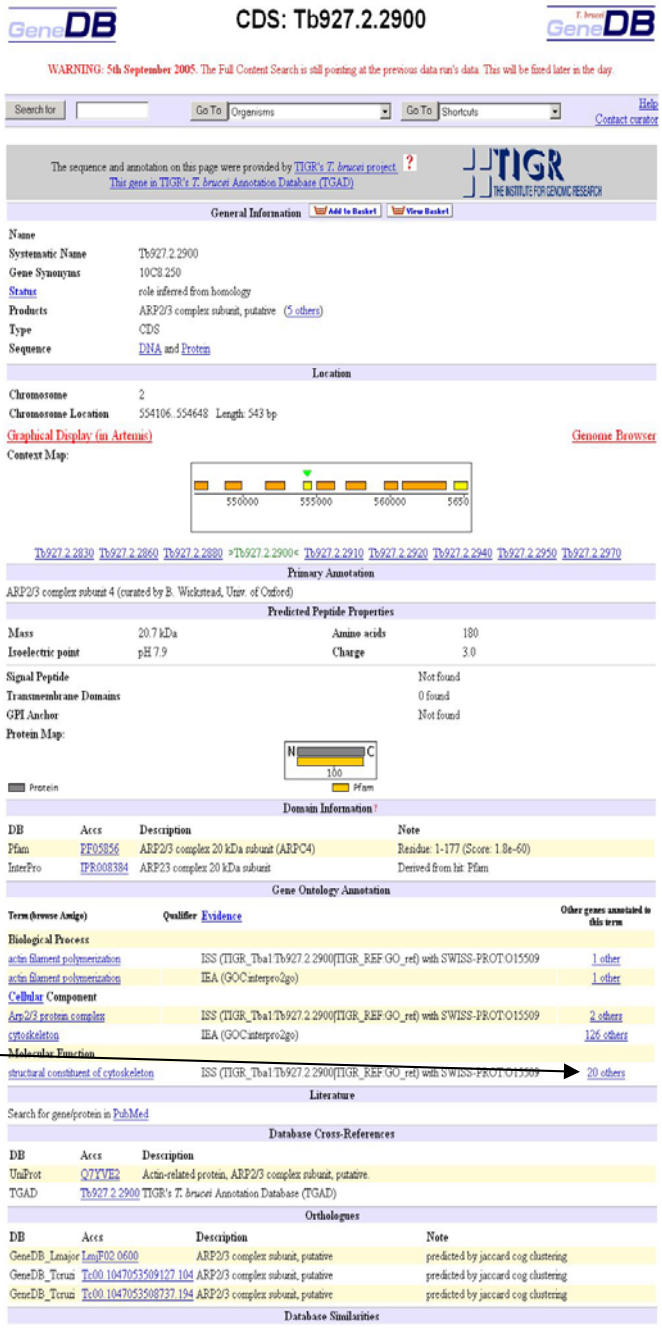
• We're going to use GO and similarity searching to identify those last two remaining components.

Gene Ontologies are structured vocabularies that are designed to describe biological processes in an accurate and consistent way (for more information see http://www.geneontology.org.) It is composed of three separate ontologies, describing aspects of a given protein's function in terms of its molecular function, biological process and cellular component (location). Where evidence exists from the literature, from sequence analysis or other sources gene ontology terms for function, process and component are attributed to that gene. AmiGO is the database housing assigned gene ontology associations and is maintained by the Gene Ontology consortium. It allows searching and browsing of gene ontology annotations across many genomes from human, mouse through to lower eukaryotes (including those which are not annotated and curated for GeneDB). GeneDB has a copy of the GO database and an installation of the AmiGO browser on top of it. Advantages of a local copy of the GO database include an increased update frequency as well as the inclusion of datasets not otherwise searchable via the 'official' GO database (e.g. assignments inferred by eletronic annotation).

It can be a powerful way to search for genes with similar function across several organisms. The example below shows how to set up this query, which can be either accessed from the organism home page and/or the search menu bar at the top of each of the feature pages. One you've tried it and have become familiar with it, try some of the other suggested searches or perhaps one that would be of interest to your own research.

1

Go to the feature page for Tb927.2.2900, one of the candidates you've come across earlier. You can type the gene name straight into the search box at the top of a feature page.

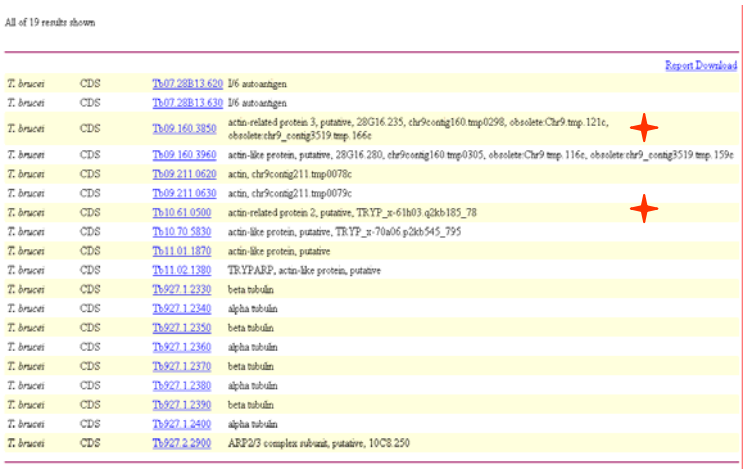


2

Arp2 can be classified as a structural constituent of the cytoskeleton. Click on the link to other genes annotated to the GO term to see which other 20 proteins have been annotated to this GO term.

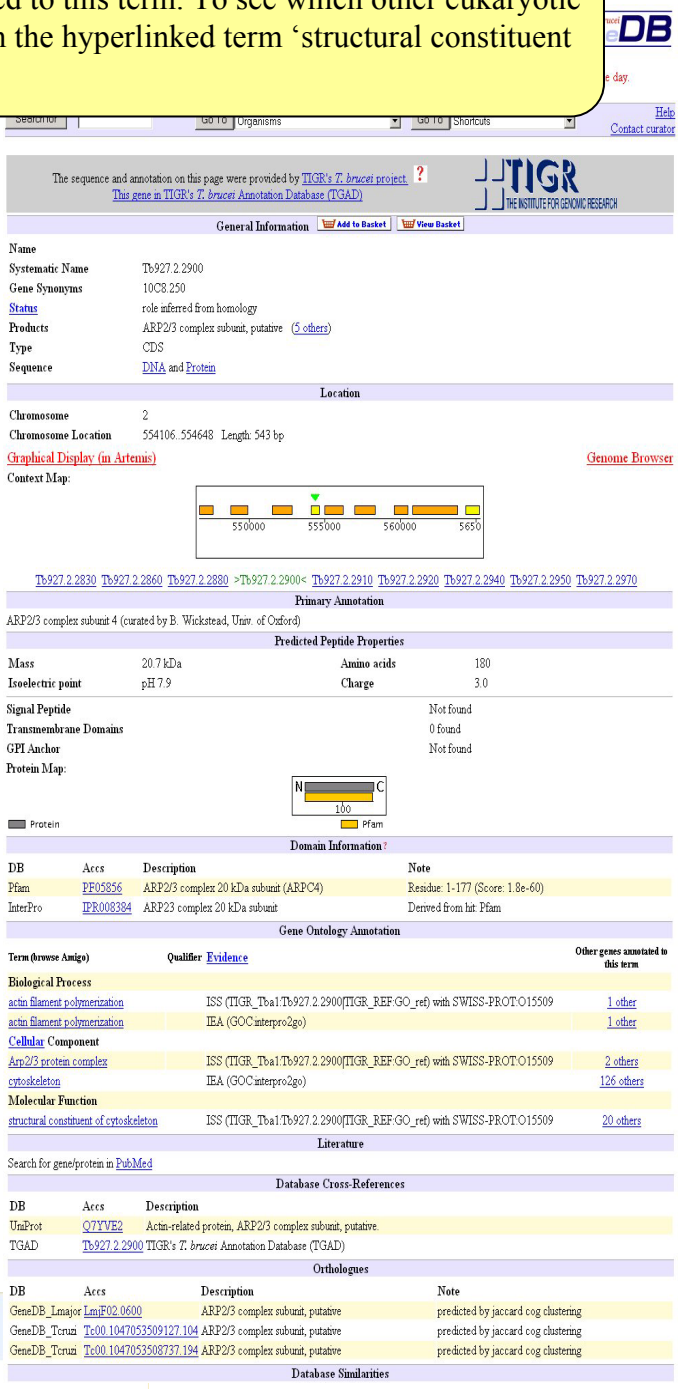
3

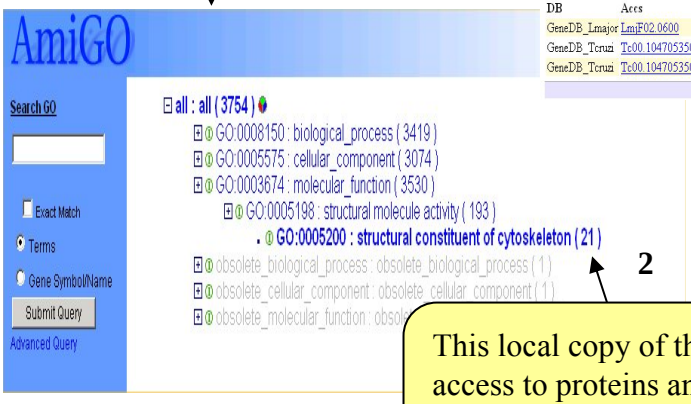
Can you identify the two missing actin-related proteins Arp2 and Arp3 from this list?



Exercise 1.4 Exploring the GeneDB AmiGO browser further

1 The local AmiGO browser not only provides access to *T. brucei* terms annotated to them same but also to proteins from other organisms annotated to this term. To see which other eukaryotic proteins have been annotated to this term, click on the hyperlinked term 'structural constituent of the cytoskeleton'.

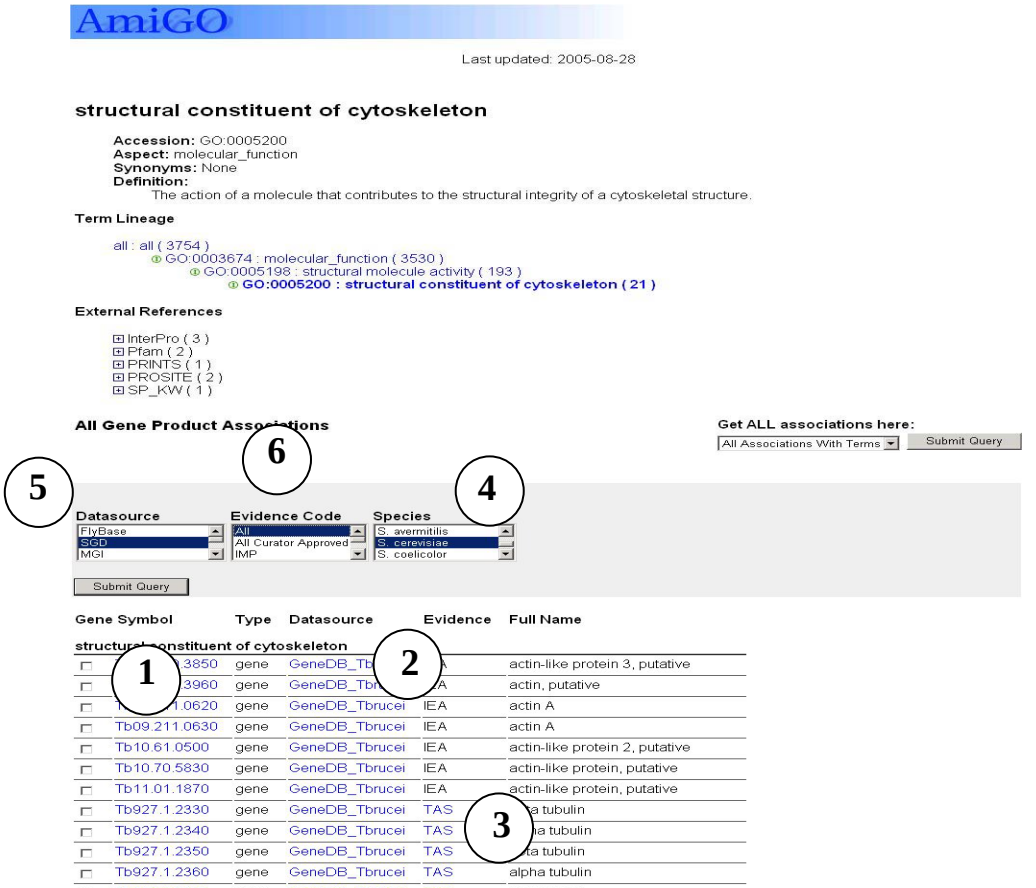




2 This local copy of the AmiGO browser provides access to proteins annotated to the same ontology term. Initially, you will see only the *T. brucei* proteins annotated to this term. Click on the hyperlinked term.

3

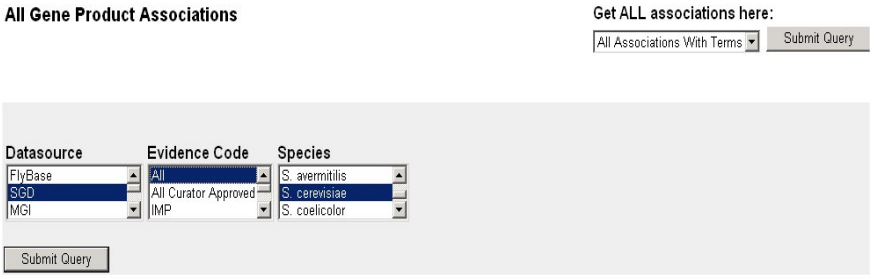
This is the top of the browser page showing the descriptions of the *T. brucei* proteins. Clicking on **1** will show you the other ontologies this particular protein is annotated to. **2** will take you to the feature page in GeneDB and the evidence cofe **TAS** (**3**) will take you to the paper substantiating this assignment.



By changing the filter settings, you can retrieve proteins from other organisms annotated to this term as well either by species name (**4**)as or database (**5**) restricting the search to a term associated with a product via a particular evidence code (**6**). Try this by searching for cytoskeleton constituents in the yeast *S. cerevisiae*. The *S. cerevisiae* database is called SGD.

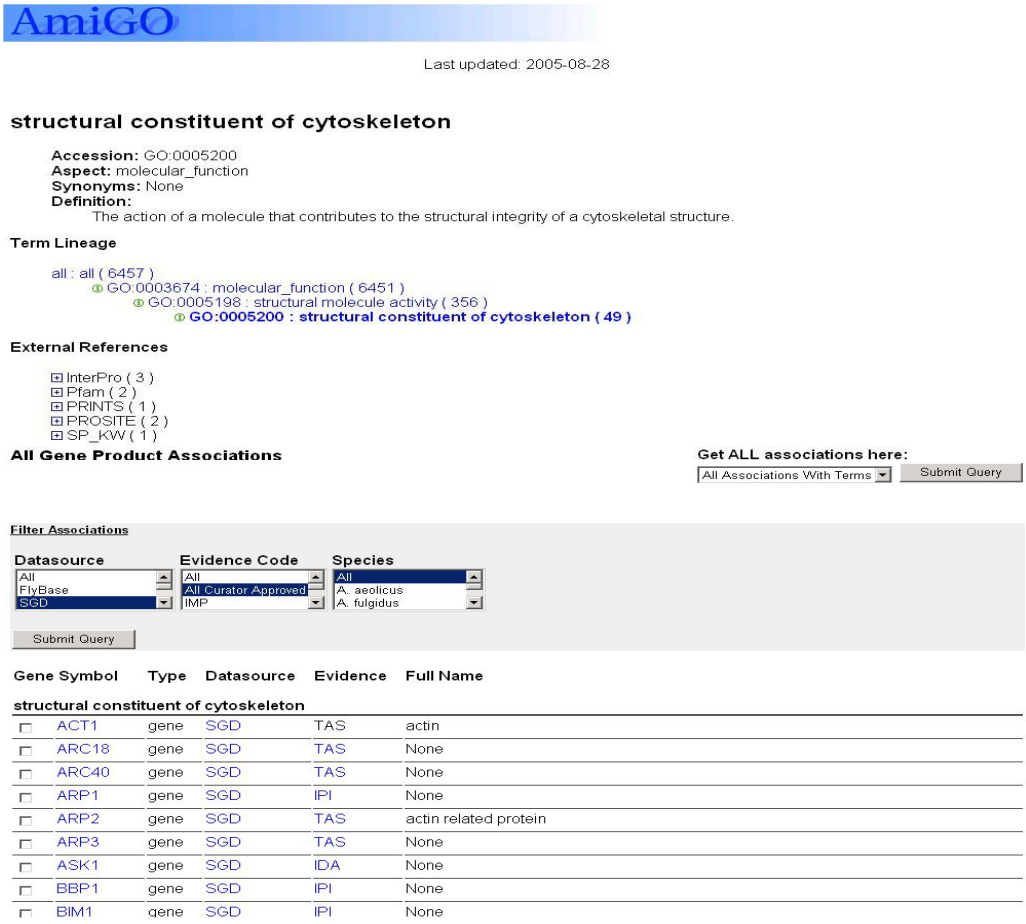
4

Select '**SGD**' and/or *S cerevisiae* from the pull down menu and press '**Submit Query**'.



5

Below are some of the proteins annotated as cytoskeletal components in *S. cerevisiae*. Click on the link to SGD and have a look at the status of annotation of this particular protein. Also, explore what happens if you were to filter on some of the other GO evidence codes such as **IDA**.



AmiGO

Last updated: 2005-08-28

structural constituent of cytoskeleton

Accession: GO:0005200
Aspect: molecular_function
Synonyms: None
Definition:
The action of a molecule that contributes to the structural integrity of a cytoskeletal structure.

Term Lineage

all : all (6457)
GO:0003674 : molecular_function (6451)
GO:0005198 : structural molecule activity (356)
GO:0005200 : structural constituent of cytoskeleton (49)

External References

InterPro (3)
Pfam (2)
PRINTS (1)
PROSITE (2)
SP_KW (1)

All Gene Product Associations

Get ALL associations here: All Associations With Terms Submit Query

Filter Associations

Datasource: All, FlyBase, SGD
Evidence Code: All, All Curator Approved, IMP
Species: All, A. aeolicus, A. fulgidus

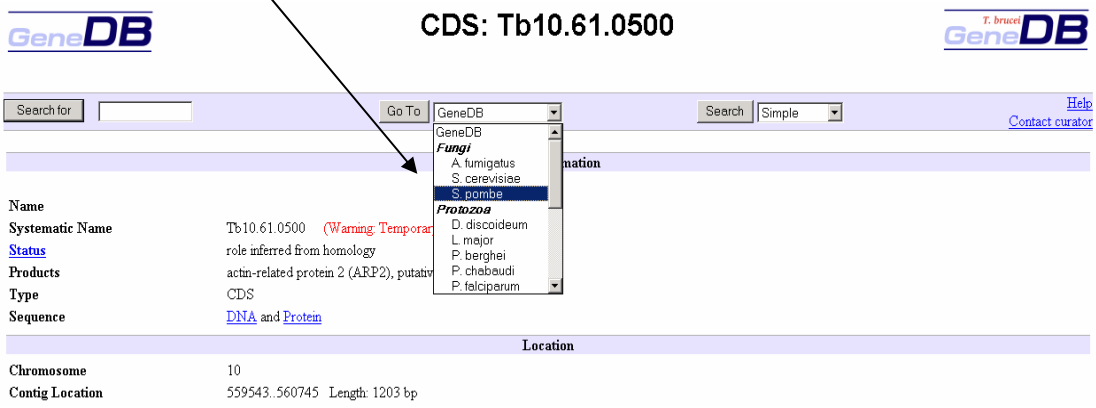
Submit Query

Gene Symbol	Type	Datasource	Evidence	Full Name
structural constituent of cytoskeleton				
ACT1	gene	SGD	TAS	actin
ARC18	gene	SGD	TAS	None
ARC40	gene	SGD	TAS	None
ARP1	gene	SGD	IPI	None
ARP2	gene	SGD	TAS	actin related protein
ARP3	gene	SGD	TAS	None
ASK1	gene	SGD	IDA	None
BBP1	gene	SGD	IPI	None
BIM1	gene	SGD	IPI	None

Exercise 1.5 The use of BLAST to identify sequences with similarity to known Arp2/3 complex subunits

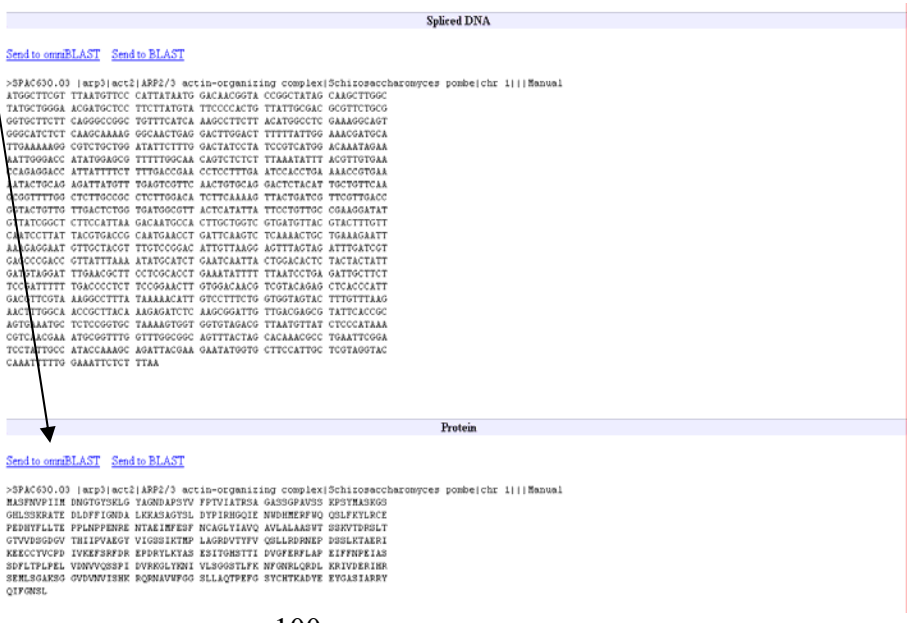
In addition to using gene ontology assignments, you could have also used similarity searches to identify the two missing actin-related proteins. To identify the putative *T. brucei* Arp3 homologue, you could also use the BLAST tool integrated into GeneDB.
As mentioned previously, the components of the Arp2/3 complex are highly conserved amongst eukaryotes. Therefore, one approach to identifing putative homologues in *T. brucei* is to use the *S. pombe* Arp3 sequence as a query sequence in a BLAST search.

1 Go to the *S. pombe* Arp3 feature page. You can do this by going to the *S. pombe* homepage via the navigation bar at the top of the any gene page (e.g.) Tb10.61.0500 and searching for '**Arp3**'.



2 Retrieve the amino acid sequence by clicking on the sequence link.

3 Click on '**send to omniBLAST**' link. omniBLAST permits searching across different databases selected by the user.

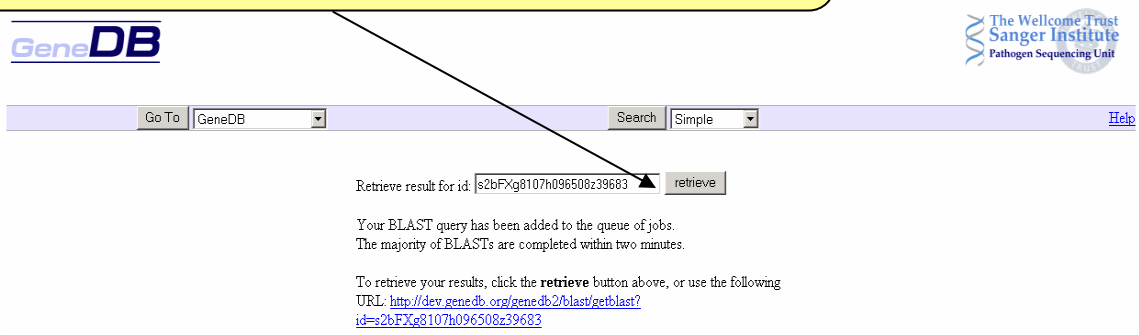


4

This is the omniBLAST page, providing access to databases of all sequences housed in GeneDB. By default, the datasets of the organism you started out from will be selected. For this search select the *S. pombe* and *T. brucei* protein databases. Note that the amino acid sequence of the *S. pombe* Arp3 protein has automatically been pasted into the query sequence box.
Start omniBLAST by clicking on the '**Start omniBLAST'** button.

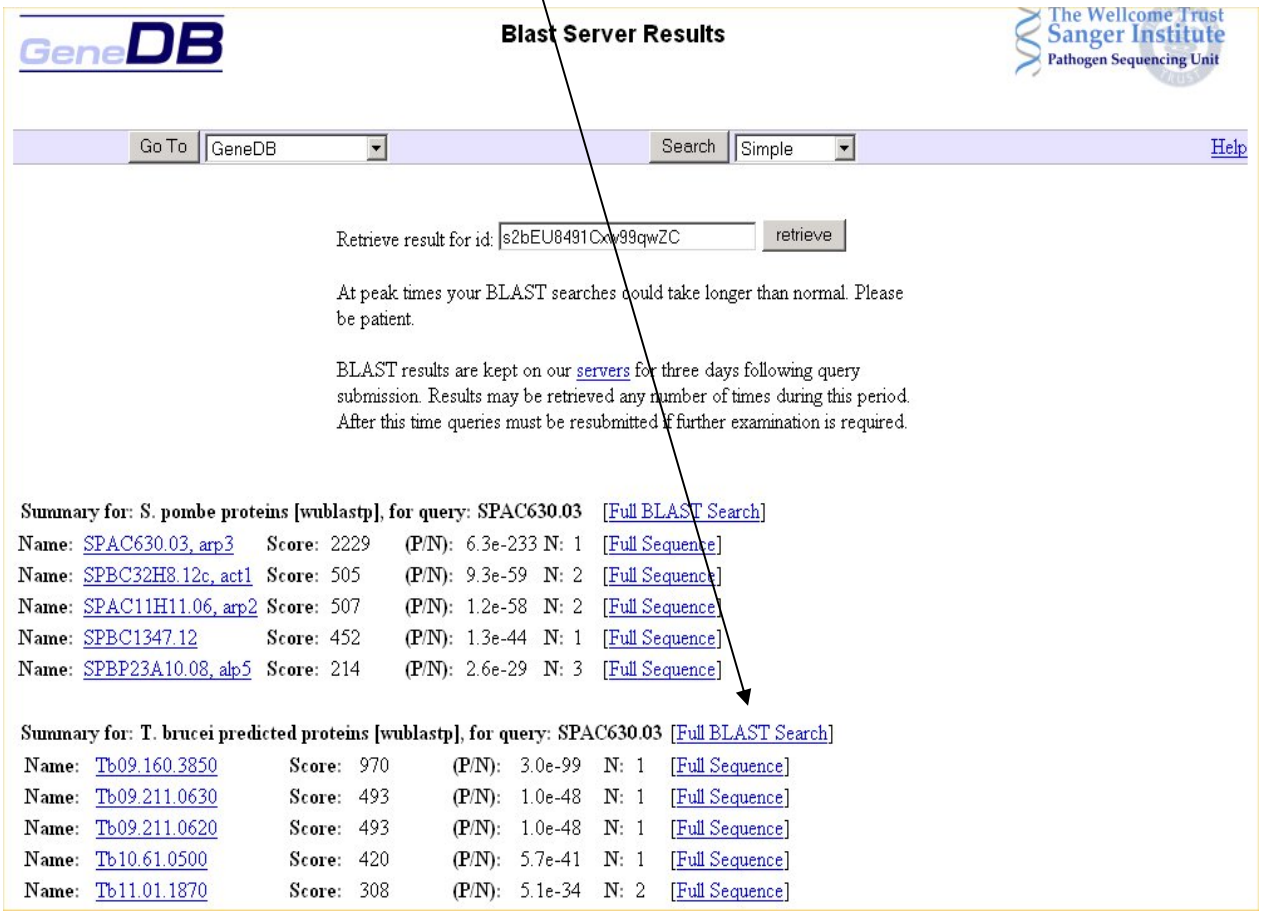
5

Retrieve the results by clicking on the '**retrieve**' button. As indicated, results will be accessible for the next 2 weeks using the listed URL.



6

You will retrieve an abbreviated BLAST results page, listing only the top 5 hits without alignments. Click on the '**Full BLAST Search**' of your query sequence against the predicted *T. brucei* proteome.



GeneDB

Blast Server Results

The Wellcome Trust Sanger Institute
Pathogen Sequencing Unit

Go To | GeneDB | Search | Simple | Help

Retrieve result for id: s2bEU8491Cxw99qwZC retrieve

At peak times your BLAST searches could take longer than normal. Please be patient.

BLAST results are kept on our servers for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

Summary for: S. pombe proteins [wublastp], for query: SPAC630.03 [Full BLAST Search]

Name:	SPAC630.03, arp3	**Score:** 2229	**(P/N):** 6.3e-233	**N:** 1	[Full Sequence]
Name:	SPBC32H8.12c, act1	**Score:** 505	**(P/N):** 9.3e-59	**N:** 2	[Full Sequence]
Name:	SPAC11H11.06, arp2	**Score:** 507	**(P/N):** 1.2e-58	**N:** 2	[Full Sequence]
Name:	SPBC1347.12	**Score:** 452	**(P/N):** 1.3e-44	**N:** 1	[Full Sequence]
Name:	SPBP23A10.08, alp5	**Score:** 214	**(P/N):** 2.6e-29	**N:** 3	[Full Sequence]

Summary for: T. brucei predicted proteins [wublastp], for query: SPAC630.03 [Full BLAST Search]

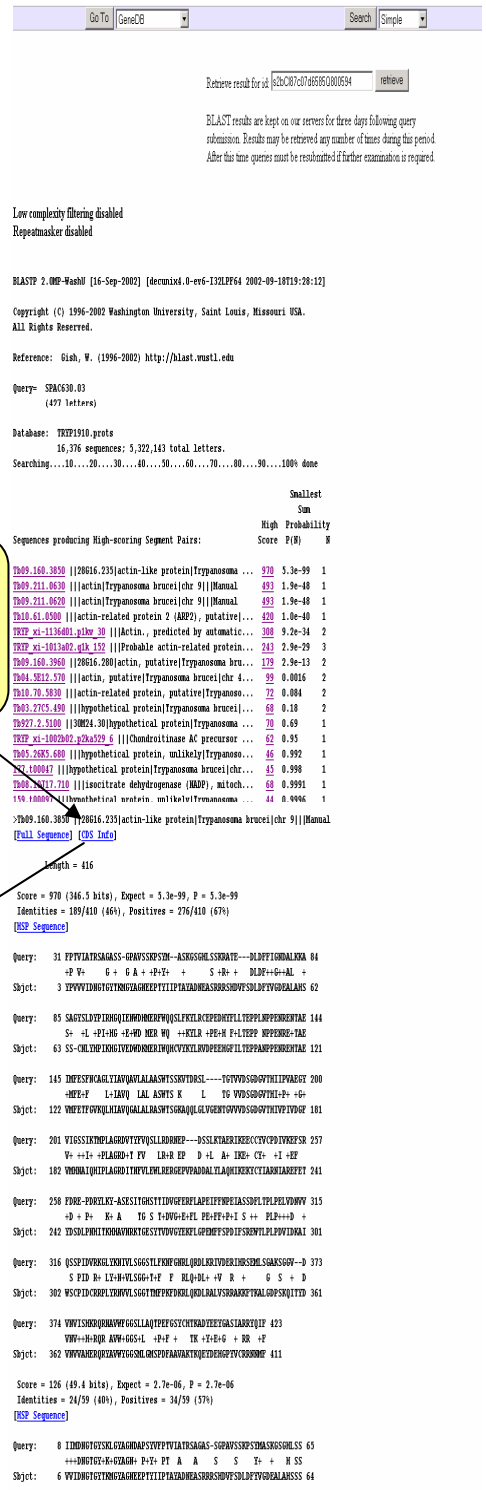
Name:	Tb09.160.3850	**Score:** 970	**(P/N):** 3.0e-99	**N:** 1	[Full Sequence]
Name:	Tb09.211.0630	**Score:** 493	**(P/N):** 1.0e-48	**N:** 1	[Full Sequence]
Name:	Tb09.211.0620	**Score:** 493	**(P/N):** 1.0e-48	**N:** 1	[Full Sequence]
Name:	Tb10.61.0500	**Score:** 420	**(P/N):** 5.7e-41	**N:** 1	[Full Sequence]
Name:	Tb11.01.1870	**Score:** 308	**(P/N):** 5.1e-34	**N:** 2	[Full Sequence]

7

This is the BLAST results page, providing information on the score, percentage identity and the alignments.

8

Evaluate the alignments



9

Clicking on '**CDS Info**' will take you to the feature page of the putative *T. brucei* Arp3 homologue.

10

	S. pombe	***T. brucei***	
Arp2	SPAC11H11.06	Tb10.61.0500	
Arp3	SPAC630.03		
p41-Arc	SPBC14C8.06		
p34-Arc (PF04045)	SPAC6F6.10c		
p21-Arc (PF04062)	SPBC1778.08c	Tb10.70.2680	
p20-Arc (PF05856)	SPAC6G9.07c	Tb927.2.2900	
p16-Arc (PF04699)	SPAC17G8.04c		

Using 4 different approaches to retrieve/identify putative homologues, you should have completed this table. As you will have noticed, you probably wouldn't have been able to retrieve all the data by just using a single approach to mine the *T. brucei* genome, which highlights some of the issues outlined in the introduction to this module.

Exercise 1.6 Identify the Arp2/3 complex in other kinetoplastid species

Imagine now that you are not only interested in this complex in *T. brucei* but also in other *Trypanosoma* and *Leishmania*, causative agents of Leishmaniasis, species. GeneDB is ideally suited to this purpose as it houses sequence and annotation of multiple kinetoplastid species and the data are extensively cross-linked. You could start by identifying components in *L. major* and *T. cruzi* and then move on to *l. infantum* and the cattle-infective *T. vivax* and *T. congolense*.

There are a number of ways you could tackle this problem. You could use similarity searches, GO and/or Pfam catalogues, similar to what you have been doing in the previous exercises. However, a faster way would be to make use of orthologue cross-links provided by GeneDB.

1

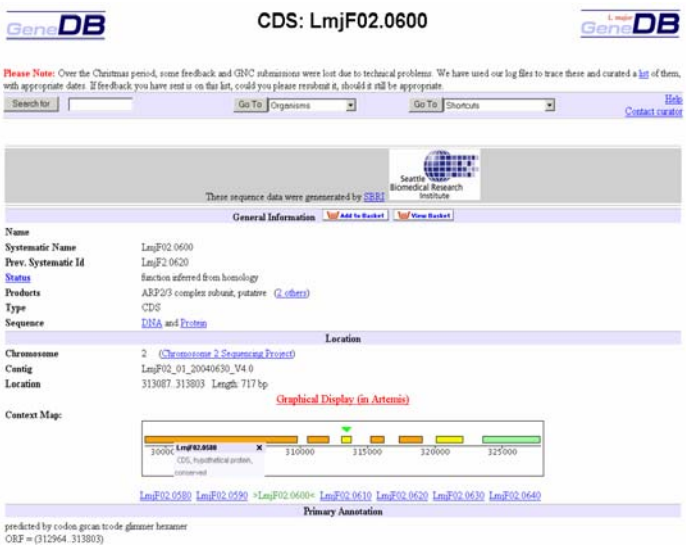
There are two ways you can do this, either by using the orthologue cross-links provided on each of the gene pages or alternatively, use the '**List Download'** option.
Start with having a look at the orthologue cross-links on the gene page. Go to the gene page showing annotation associated with Tb927.2.2900.

2

This part of the gene page provides links to manually curated orthologues in other species.

3

Click on the link providing access to the putative *L. major* orthologue.



4

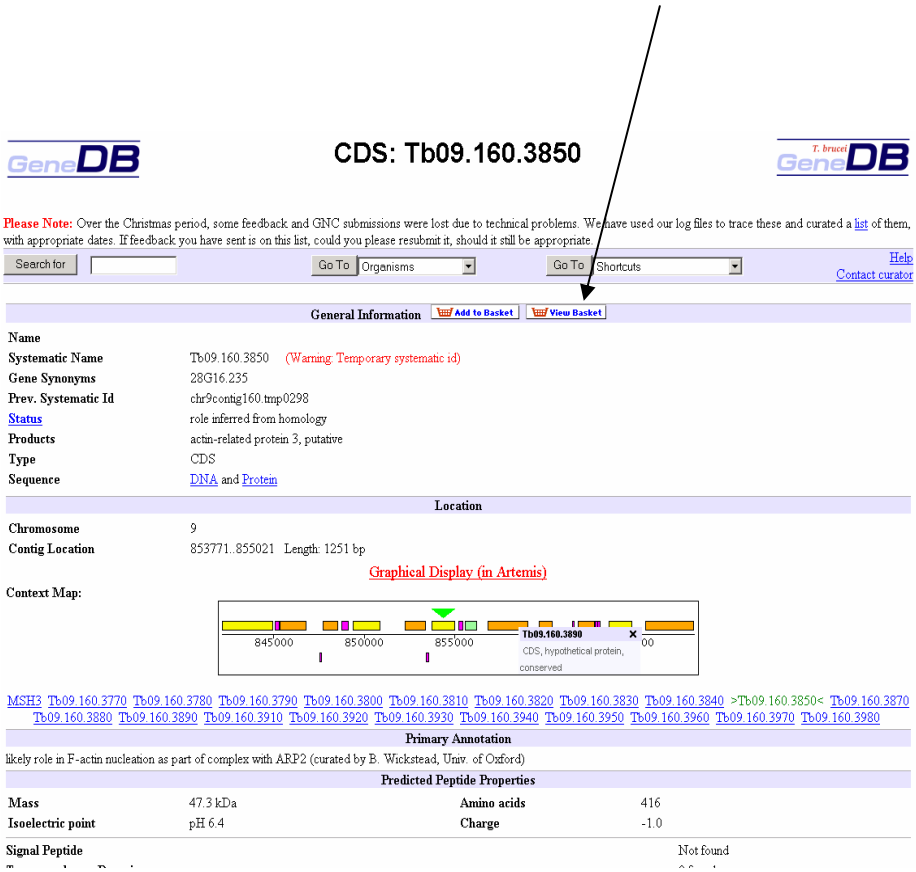
You could now go through each one of the 7 putative members of the *T. brucei* Arp2/3 complex, identifying putative orthologues by looking at the '**Orthologues**' section on the gene page. There is however a faster way using the '**List Download**' utility. This function allows you to compile a list of your gene of interest and then subsequently downloading the description and sequence of these features using the **'Gene Basket**'.
Start by going to the top of the gene page. Imagine this to be your first gene of interest. In order collect your genes of interest, you'll need to click on the '**Add to Basket**' icon at the top of the page. This will now have added the identifier of this gene to the virtual basket.

5

Now go to each one of the gene pages of the putative Arp2/3 complex members – they are:
Tb10.70.2680
Tb10.61.0500
Tb10.389.0270
Tb10.406.0320
Tb927.2.2900
Tb927.8.4410
Tb09.160.3850
and should all be listed in the table you filled in earlier. You can simply navigate between gene pages by filling in the '**Search for**' box in the navigation bar. Add them to the gene basket by simply clicking on the '**Add to Basket**' icon at the top of each of the gene pages.

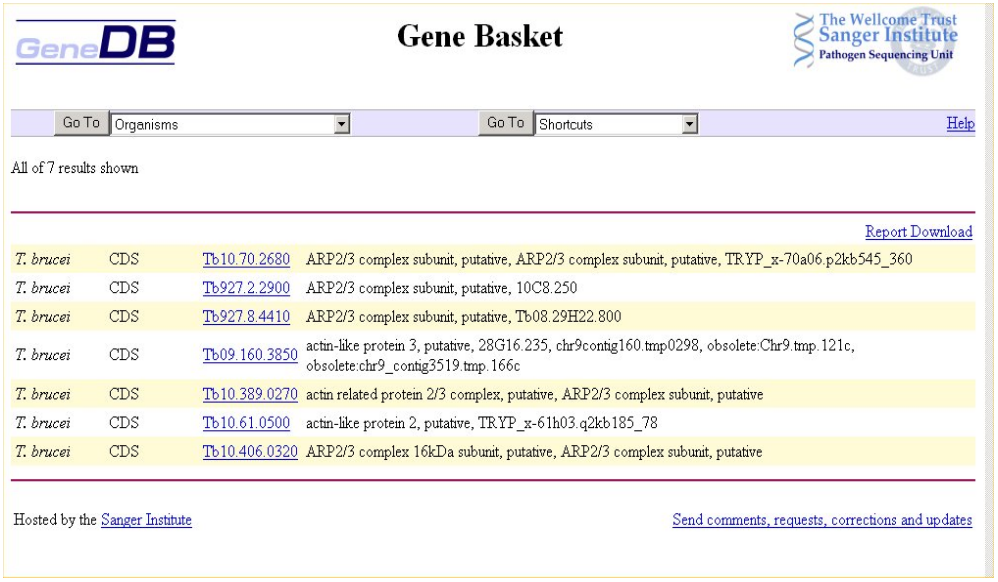
6

Once you have added all your gene of interest to the basket, click on the '**View Basket**' icon.



7

The page lists all your genes of interest in the gene basket.

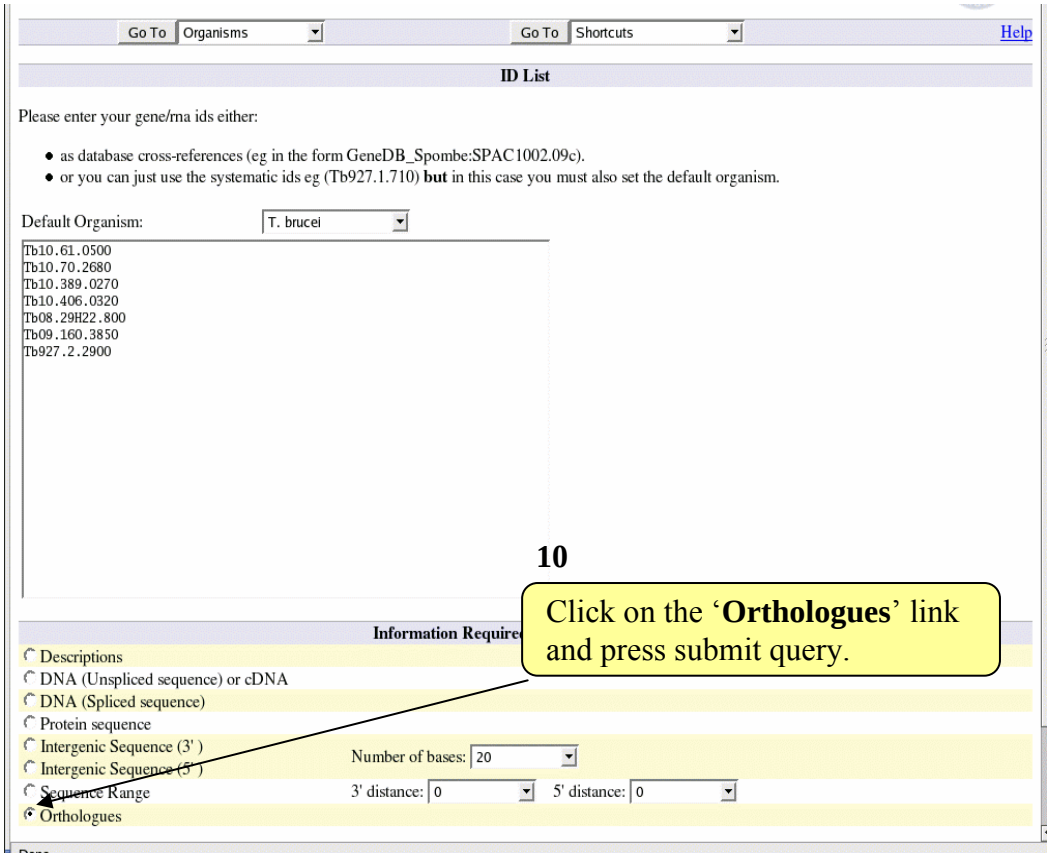


8

Click on the '**Report Download**' link.

9

This page allows you to download the description, sequence (both nucleotide and amino acid) and more importantly, the orthologues of your gene of interest.



10

Click on the '**Orthologues**' link and press submit query.

11

The browser window return the annotated orthologues in *L. major* and *T. cruzi*.

T. brucei	Tb09.160.3850	GeneDB_Tcruzi:TcCO.1047053508277.260, GeneDB_Tcruzi:TcCO.1047053503913.20
T. brucei	Tb927.8.4410	GeneDB_Lmajor:LmjF10.1000, GeneDB_Tcruzi:TcCO.1047053506865.10, GeneDB_Tcruzi:TcCO.104705350
T. brucei	Tb10.70.2680	GeneDB_Tcruzi:TcCO.1047053508625.30, GeneDB_Tcruzi:TcCO.1047053510963.70
T. brucei	Tb927.2.2900	GeneDB_Lmajor:LmjF02.0600, GeneDB_Tcruzi:TcCO.1047053509127.104, GeneDB_Tcruzi:TcCO.104705350
T. brucei	Tb10.406.0320	GeneDB_Tcruzi:TcCO.1047053511635.30
T. brucei	Tb10.389.0270	GeneDB_Lmajor:LmjF18.0920, GeneDB_Tcruzi:TcCO.1047053504215.40, GeneDB_Tcruzi:TcCO.104705351
T. brucei	Tb10.61.0500	GeneDB_Tcruzi:TcCO.1047053508899.110, GeneDB_Tcruzi:TcCO.1047053511361.40

12

Have a look at the results.
- Is *L. major* predicted to contain all components of the Arp2/3 complex?
- How many copies of each of the components is *T. cruzi* genome predicted to encode?
- Why might you find them in duplicates (hint: have a look at the *T. cruzi* GeneDB homepage)?
- How would you go about identifying homologues in the other *Leishmania* and *Trypanosoma* species?
- Send a feedback e-mail to the curators in cases where you come across mis-annotation.

	T. brucei	*L. major*	*T. cruzi*	*L. infantum*	*T. vivax*	*T. coongolense*	*T. b. gambiense*
Arp2	Tb10.61.0500						
Arp3	Tb09.160.3850						
p41-Arc	Tb10.380.0270						
p34-Arc (PF04045)	Tb927.8.4410						
p21-Arc (PF04062)	Tb10.70.2680						
p20-Arc (PF05856)	Tb927.2.2900						
p16-Arc (PF04699)	Tb10.406.0320						

During this exercise you will have become familiar with GeneDB, the way data are displayed on feature pages and the various ways data can be accessed. As you will have seen, you wouldn't have been able to retrieve all the data by just using a single approach to mine the genomes, but that instead you needed to employ multiple search strategies. You will also have seen how to compile lists of genes of interest and how to download them for further examination/ experimentation. Lastly, with the increasing emphasis on comparative genomics, you hopefully saw how GeneDB allows you to easily retrieve genes from related organisms.

Exercise 2 Use of the Artemis Applet

As you will have seen earlier, GeneDB has an integrated Artemis applet. You'll be using the applet to view additionally annotated sequence features as well as to download a range of sequence and features.
The *T. brucei* genome contains 3 copies of the gene encoding phosphoglycerate kinase. The 3 isozymes are differentially expressed during the life cycle of *T. brucei* and are targeted to different organelles. Does the sequence provide you with any clues why this may be?

1

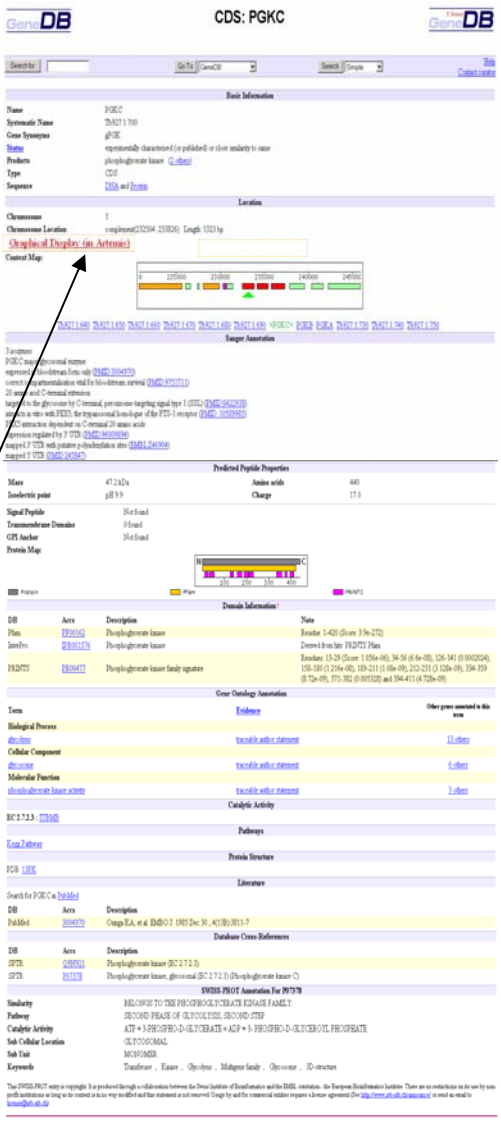
Find out where the 3 genes are located by typing '**phosphoglycerate kinase**' into the search box on the *T. brucei* homepage.

2

The results page indicates that the 3 genes are tandemly arrayed on chromosome I. Click on the the hyperlinked Tb927.1.700 gene name to get to the feature page.

3

Click on '**Graphical Display (in Artemis)**' to open up the range download page



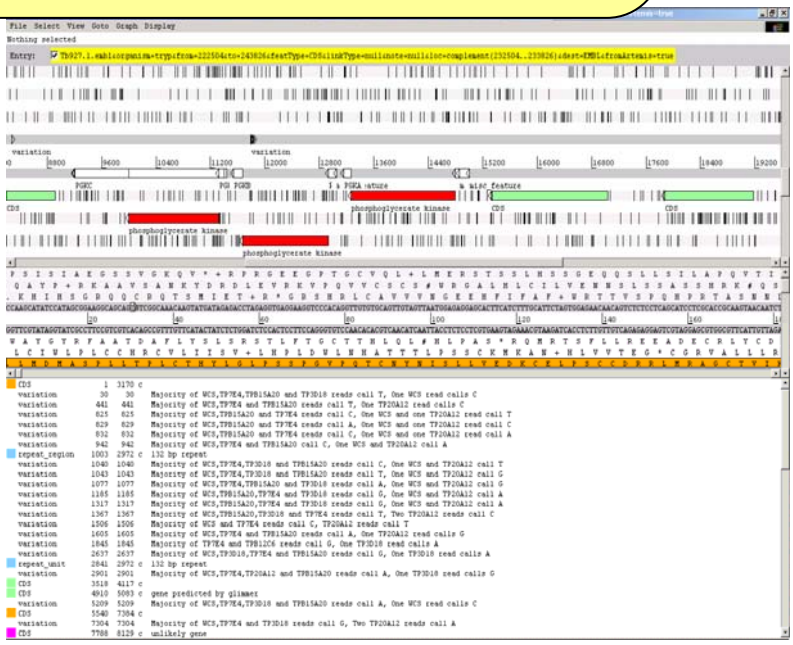
The range download page allows you to define the range of sequence you'd like to download in either EMBL or FASTA format or alternatively, open up in an Artemis applet. By default, the sequence 10kb upstream and downstream of your feature of interest will be selected.

4

Select the Artemis applet and click on the '**Submit Query**' button.

5

Find your region of interest in the applet. You'll see that in addition to the coding sequences having been annotated, the 5' and 3' UTRs have also been annotated. The 3' UTRs in particular, have been implicated in the differential regulation of mRNA stability at different life cycle stages in Kinetoplastids. In order to see whether this may also be the case here, you'll be downloading the sequence from the applet and aligning them using an alignment programme called clustalx.

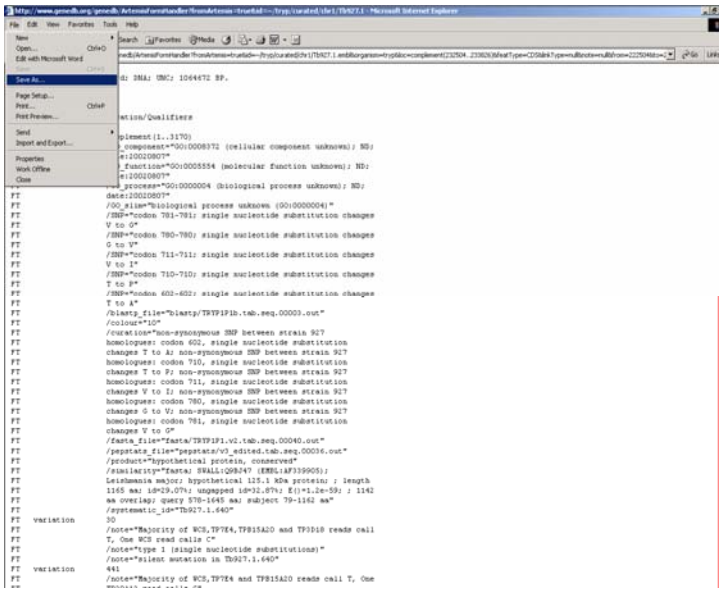


6

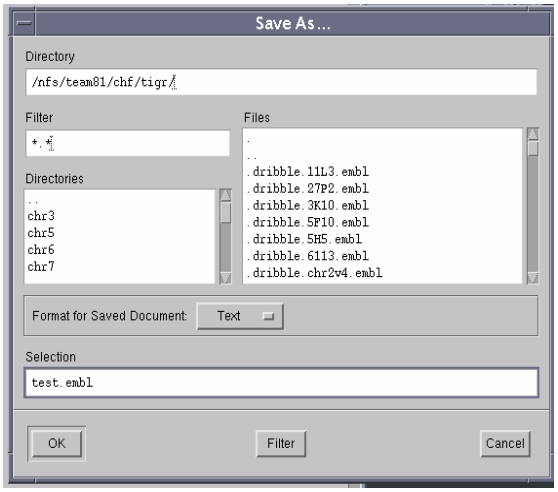
Go back to the range download page and select the option to download the sequence in EMBL format. Press the '**Submit Query**' button.

7

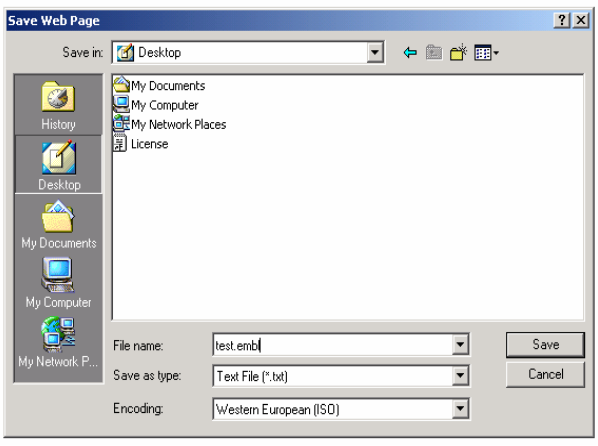
Save the sequence as a text file using your browser menu.



Unix/Linux

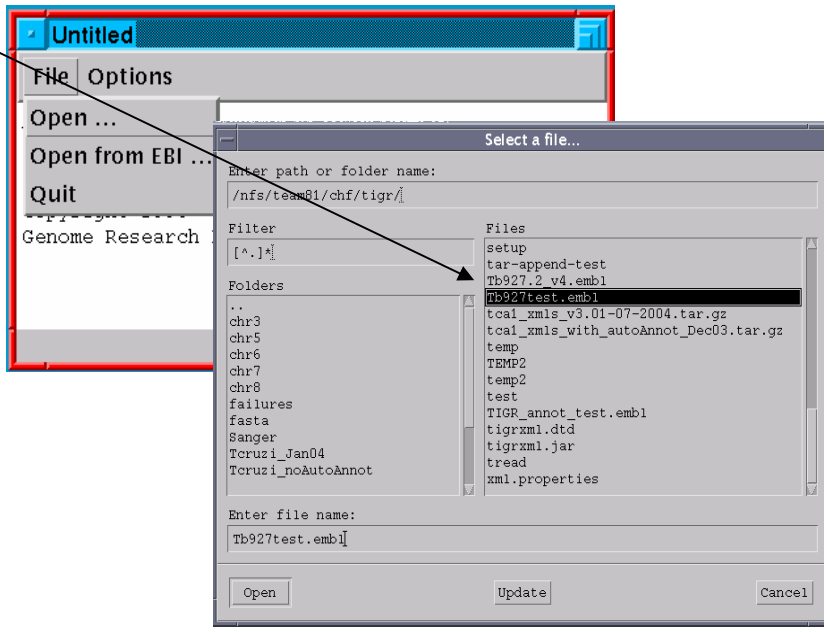


Windows



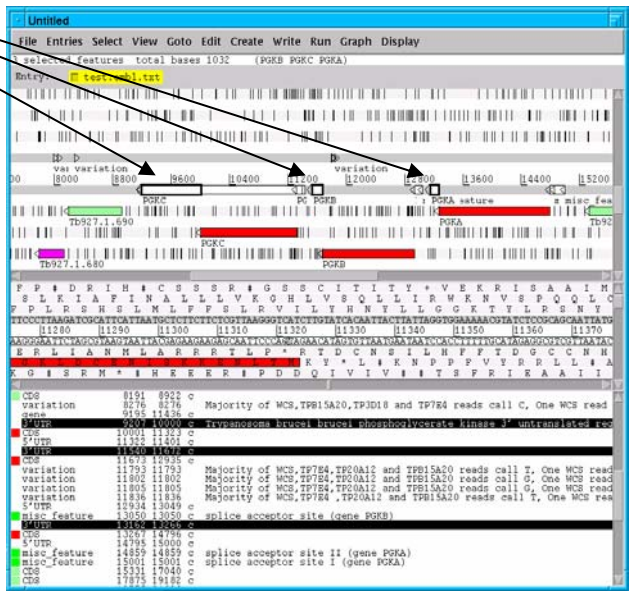
8

Open the sequence in Artemis.



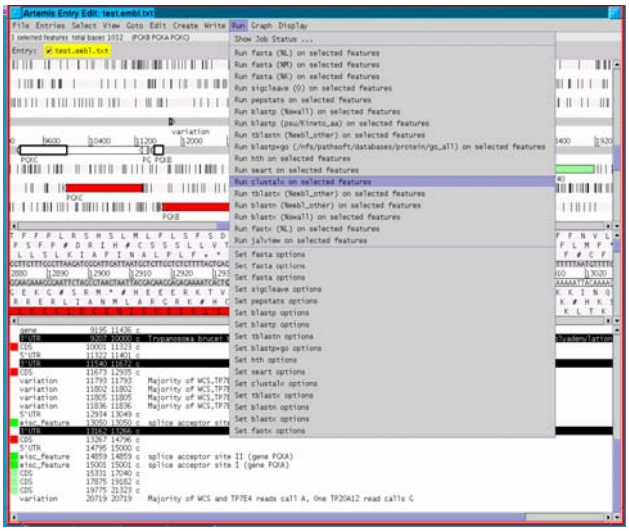
9

Select the 3' UTRs you'd like to align.



10

Select the '**Run clustalx (DNA) on selected feature**' programme from the **Run** menu.

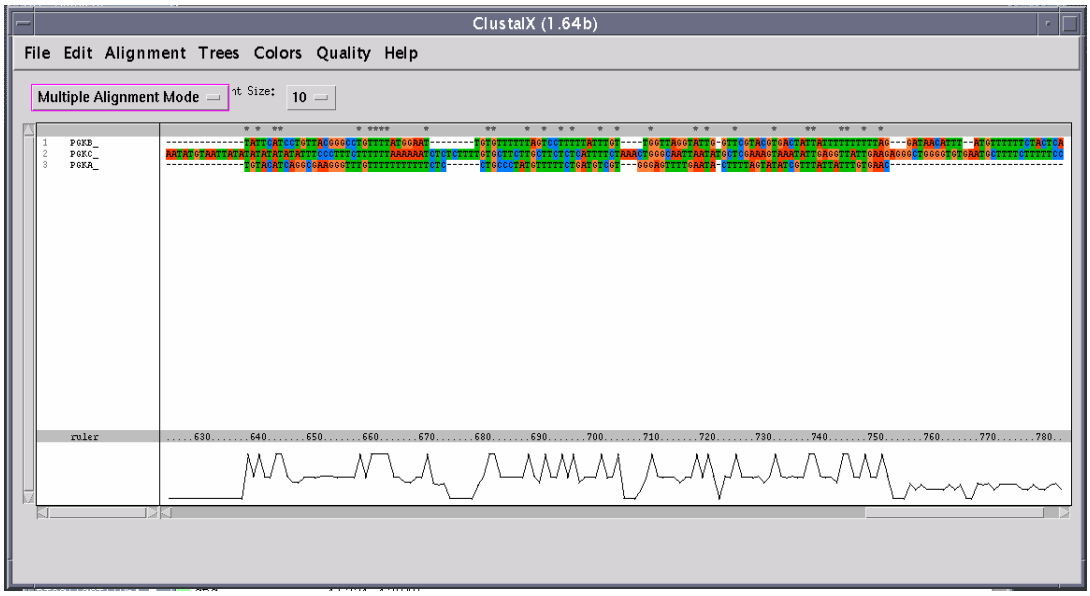


11

Change the alignment to'**Do Complete Alignment**'.

12

Look at the alignment. Do the differences in UTR length and sequence tie in with the information you can get from the GeneDB pages regarding the localisation and expression of these 3 isozymes? Have also a look at the protein sequence alignments.



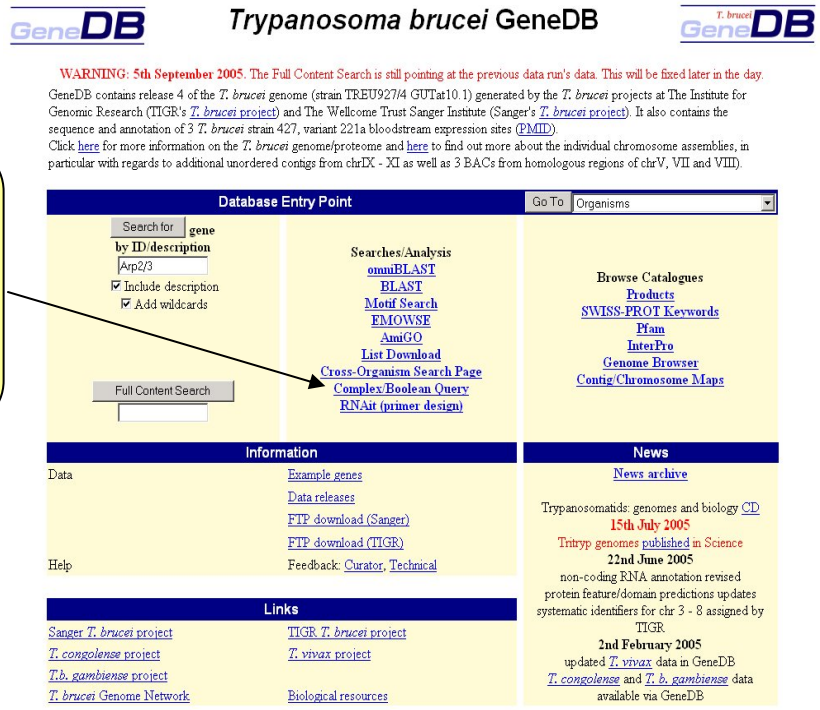
If you haven't got access to Artemis installed in the Unix/Linux environment, then you could always run the alignments using clustal via the web:
http://www.ebi.ac.uk/clustalw/

Exercise 3 Boolean querying

By now, you will have familiarised yourself with a variety of tools to search and browse the data housed in GeneDB. An additional query interface supports a wide range of queries on sequences and (curated) annotations stored in the relational database GUS. Searches can be combined with the boolean operators AND and OR. For example, users can select all proteins of a specified length range with a specified number of introns. Other query options include GO assignments, keywords, chromosome, protein domains and predicted protein sequence features. The queries in each session are tracked via a history page, allowing further refinement of searches and downloading of results as a nucleotide or amino acid FASTA file.
This exercise will demonstrate how to combine/build up queries to retrieve a subset of predicted ABC transporters containing 8 transmembrane domains.

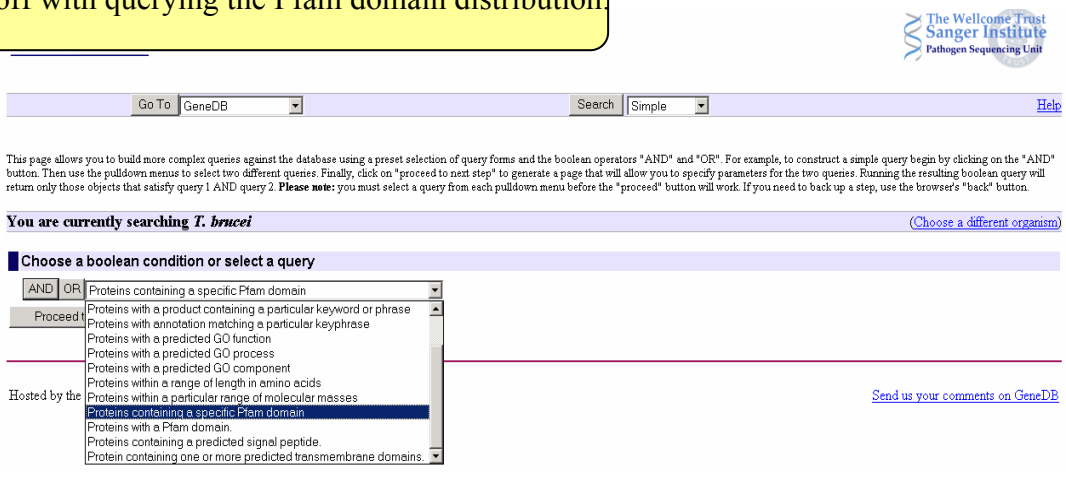
1

Select the link to the boolean querying interface by clicking on the '**complex querying pages**' button on the *T. brucei* homepage.



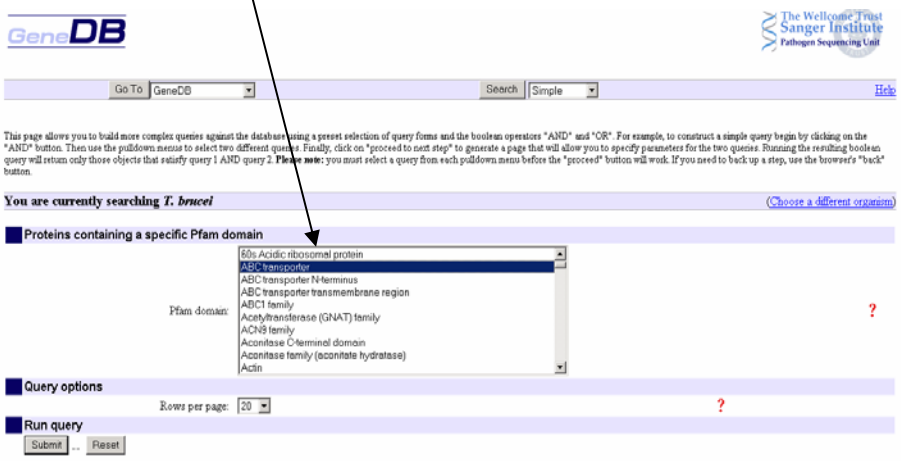
2

Start off with querying the Pfam domain distribution.



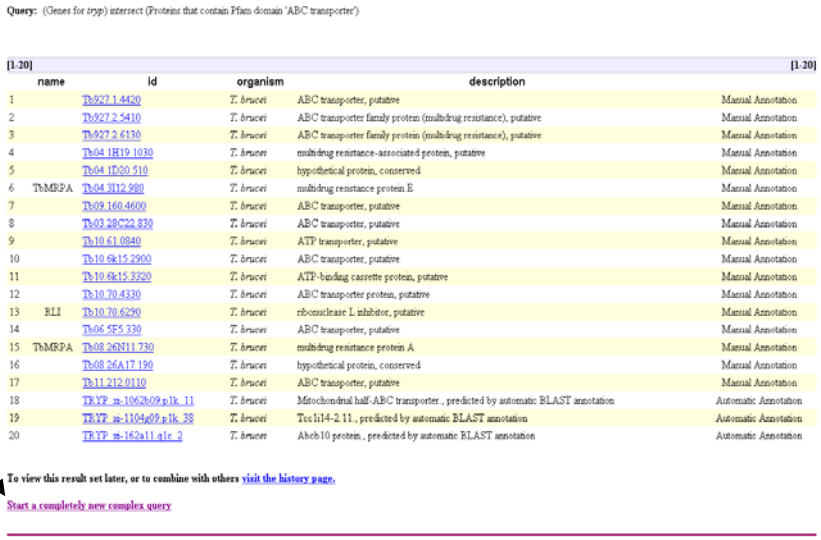
3

Select the Pfam domain **'ABC transporter**' to retrieve proteins with this predicted domain



4

This is the results list.

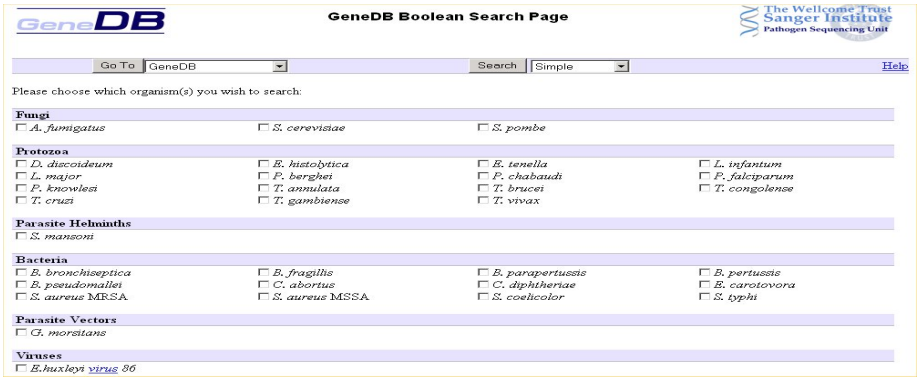


5

To identify only the predicted ABC transporters with 8 transmembrane domains, select the '**Start a completely new complex query**' from the bottom of the page.

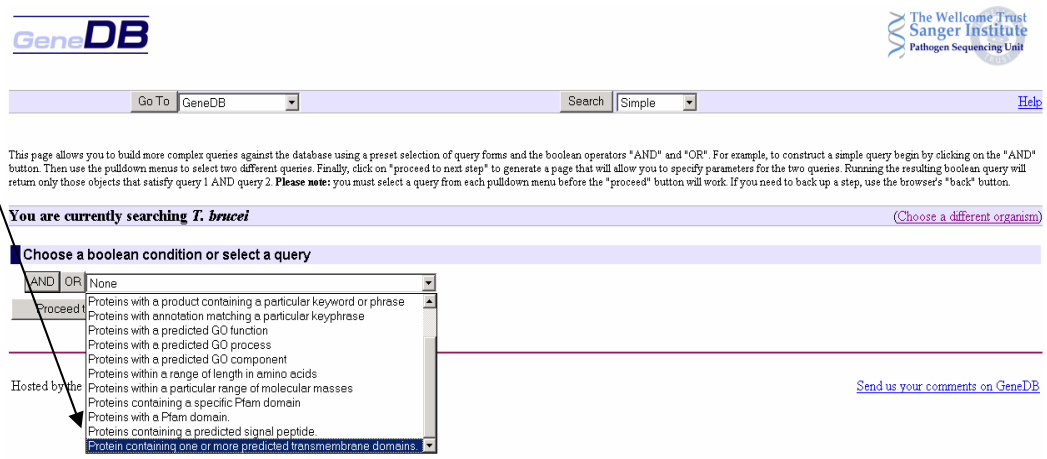
6

Select '**T. brucei**' as an organism. Note, that you can also use the boolean querying interface to search across organism datasets.



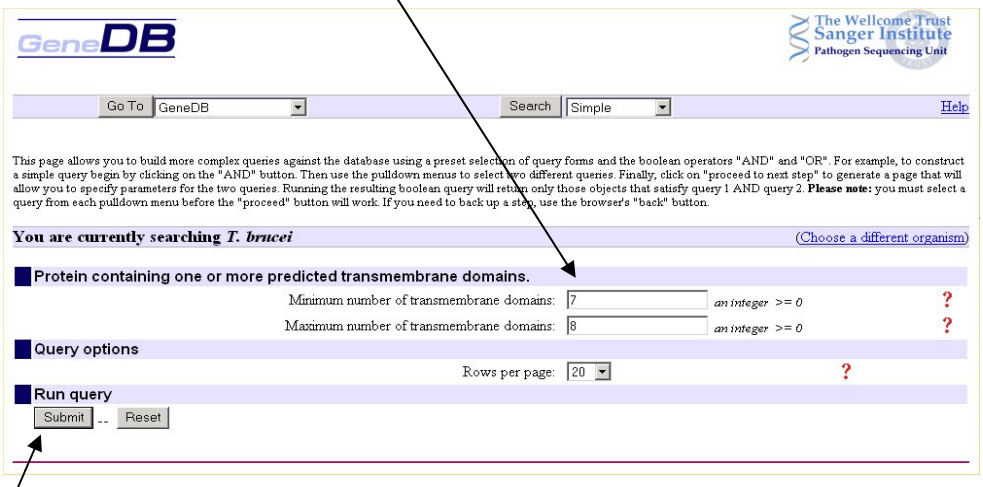
7

Select '**Protein containing one or more predicted transmembrane domains**' from the pull down menu.



8

Select the option of only retrieving proteins with 7 or 8 transmembrane domains by setting the minimum value to 7 and maximum value to 8.

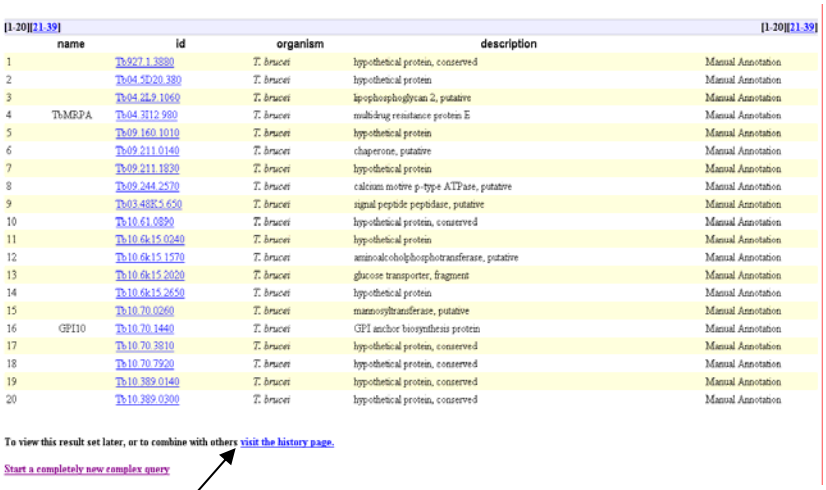


9

Press '**Submit**'

10

This is the results list of all the proteins containing 7 or 8 predicted transmembrane domains.



[1-20][21-39] [1-20][21-39]

	name	id	organism	description	
1		Tb927.1.3880	T. brucei	hypothetical protein, conserved	Manual Annotation
2		Tb04.5D20.380	T. brucei	hypothetical protein	Manual Annotation
3		Tb04.2L9.1060	T. brucei	lipophosphoglycan 2, putative	Manual Annotation
4	TbMRPA	Tb04.3I12.980	T. brucei	multidrug resistance protein E	Manual Annotation
5		Tb09.160.1010	T. brucei	hypothetical protein	Manual Annotation
6		Tb09.211.0140	T. brucei	chaperone, putative	Manual Annotation
7		Tb09.211.1830	T. brucei	hypothetical protein	Manual Annotation
8		Tb09.244.2570	T. brucei	calcium motive p-type ATPase, putative	Manual Annotation
9		Tb03.48K5.650	T. brucei	signal peptide peptidase, putative	Manual Annotation
10		Tb10.61.0890	T. brucei	hypothetical protein, conserved	Manual Annotation
11		Tb10.6k15.0240	T. brucei	hypothetical protein	Manual Annotation
12		Tb10.6k15.1570	T. brucei	aminoalcoholphosphotransferase, putative	Manual Annotation
13		Tb10.6k15.2020	T. brucei	glucose transporter, fragment	Manual Annotation
14		Tb10.6k15.2650	T. brucei	hypothetical protein	Manual Annotation
15		Tb10.70.0260	T. brucei	mannosyltransferase, putative	Manual Annotation
16	GPI10	Tb10.70.1440	T. brucei	GPI anchor biosynthesis protein	Manual Annotation
17		Tb10.70.3810	T. brucei	hypothetical protein, conserved	Manual Annotation
18		Tb10.70.7920	T. brucei	hypothetical protein, conserved	Manual Annotation
19		Tb10.389.0140	T. brucei	hypothetical protein, conserved	Manual Annotation
20		Tb10.389.0300	T. brucei	hypothetical protein, conserved	Manual Annotation

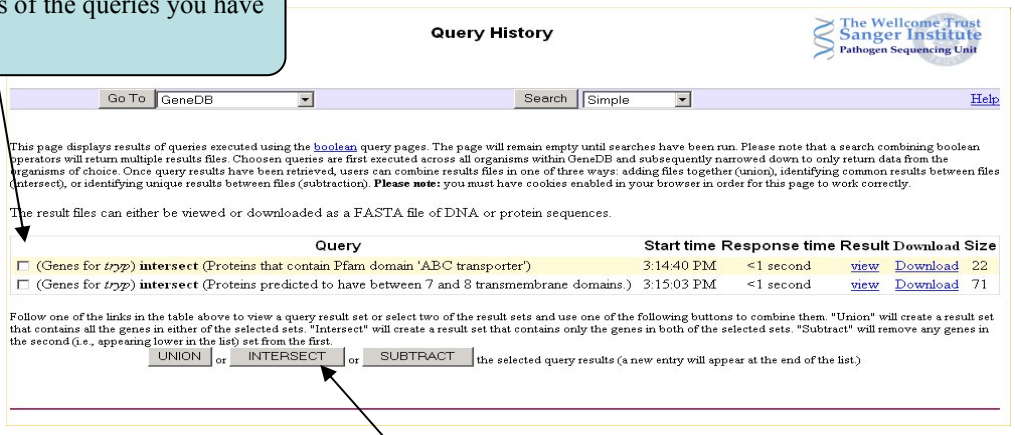
To view this result set later, or to combine with others visit the history page.

Start a completely new complex query

11

You will have now carried out 2 searches, querying GeneDB independently for the predicted ABC transporters as well as for all the proteins with 7 or 8 predicted transmembrane domains. To now identify the subset of ABC transporters with only 7 or 8 transmembrane domains, go to the history page which tracks all the queries you have executed in that session.

These are the descriptions of the queries you have executed



Query History

The Wellcome Trust Sanger Institute Pathogen Sequencing Unit

Go To GeneDB Search Simple Help

This page displays results of queries executed using the boolean query pages. The page will remain empty until searches have been run. Please note that a search combining boolean operators will return multiple results files. Choosen queries are first executed across all organisms within GeneDB and subsequently narrowed down to only return data from the organisms of choice. Once query results have been retrieved, users can combine results files in one of three ways: adding files together (union), identifying common results between files (intersect), or identifying unique results between files (subtraction). **Please note:** you must have cookies enabled in your browser in order for this page to work correctly.

The result files can either be viewed or downloaded as a FASTA file of DNA or protein sequences.

Query	Start time	Response time	Result	Download	Size
☐ (Genes for *tryp*) **intersect** (Proteins that contain Pfam domain 'ABC transporter')	3:14:40 PM	<1 second	view	Download	22
☐ (Genes for *tryp*) **intersect** (Proteins predicted to have between 7 and 8 transmembrane domains.)	3:15:03 PM	<1 second	view	Download	71

Follow one of the links in the table above to view a query result set or select two of the result sets and use one of the following buttons to combine them. "Union" will create a result set that contains all the genes in either of the selected sets. "Intersect" will create a result set that contains only the genes in both of the selected sets. "Subtract" will remove any genes in the second (i.e., appearing lower in the list) set from the first.

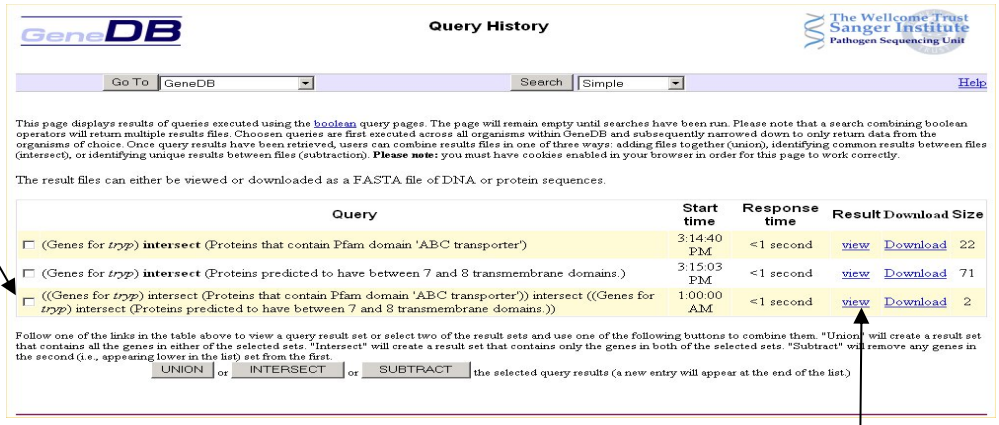
UNION or INTERSECT or SUBTRACT the selected query results (a new entry will appear at the end of the list.)

12

Via the query history page, you can view and download the sequences of your queries for further examination/manipulation. You can also combine results files and/or identify share/unique results between 2 queries. In order to identify the subset of ABC transporters, select both your queries and select '**Intersect**' button.

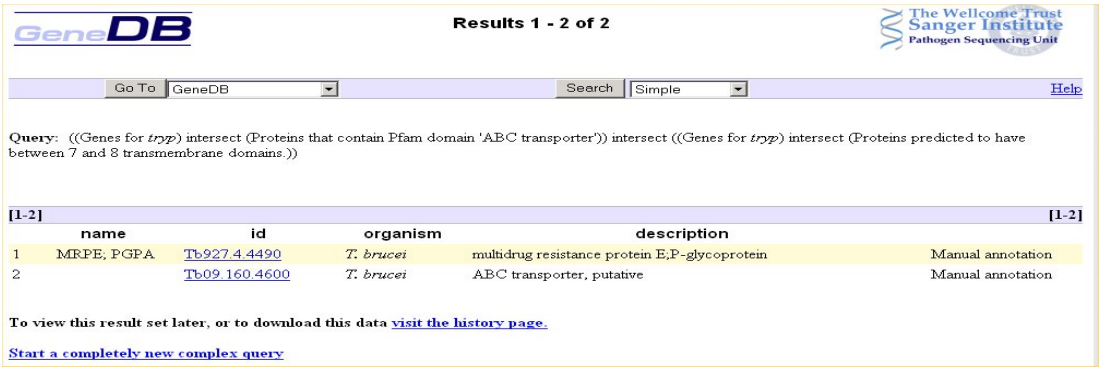
13

This will return a third set of query results which should only contain the subset of ABC transporters.

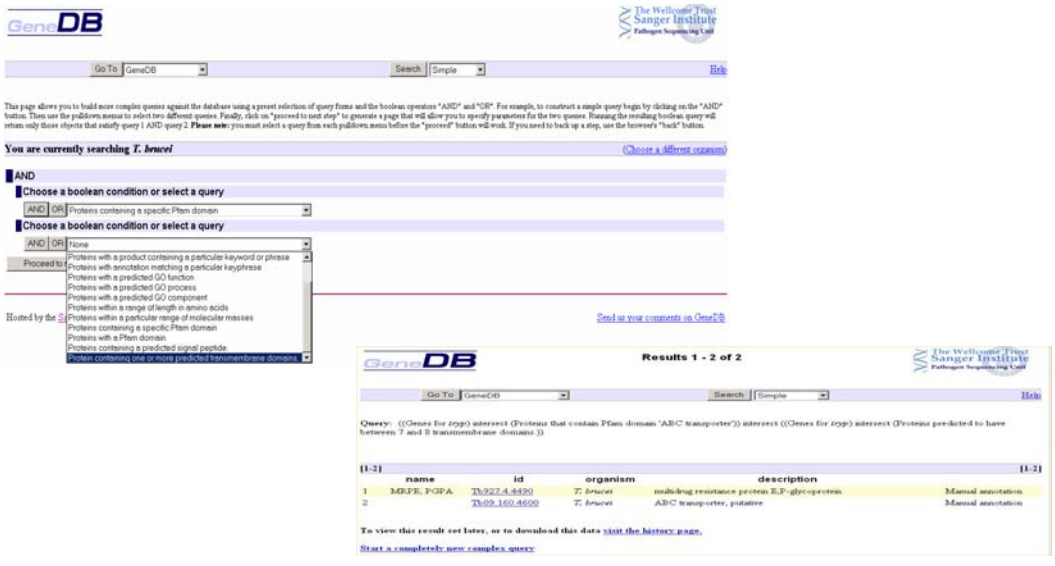


14

Click on '**view**' to have a look at the results. Clicking on the hyperlinked gene names will take you to the feature pages for these putative ABC transporters.



You would have been able to retrieve the same result by combining the two queries from the outset using '**AND**'. By clicking onto '**AND**' on the initial *T. brucei* query page, you will get the option of executing multiple queries simultaneously (see below).



Exercise 5 Identifying autotransporters encoded in the genomes of *Bordetella pertussis, B. parapertussis* and *B. bronchiseptica*.

B. pertussis, *B. parapertussis* and *B. bronchiseptica* are closely related Gram-negative β-proteobacteria. They colonize the respiratory tract of mammals, causing whooping cough (*B. pertussis*, *B. parapertussis*) as well as a chronic respiratory infection in a range of mammals (*B. bronchiseptica*).

This exercise is designed to identify autotransporters in the three genomes of the *Bordetella* spp. Autotransporters are members of a large family of exported proteins, encoding an integral outer-membrane pore which enables the bacteria to cross the outer membrane. As such, autotransporters are postulated to function in host interaction and virulence, some of which have been experimentally confirmed.

Imagine you came across a recent paper describing the autotransporter complement in *B. bronchiseptica* (see table on page 56).
Now think of ways you could identify autotransporters in the other *Bordetella* species. You could do this in a variety of ways:

• keyword searches of assigned product names (**exercise 4.1**).

• using the orthologue links provided on the gene pages (**exercise 4.2**).

• using BLAST (**exercise 4.3**).

• using Pfam/Reily browsable catalogues (**exercise 4.4**).

• using boolean querying tool (**exercise 4.5**)

Once you've identified the autotransporters across the three species, we're going to examine the genomic loci of one of these transporters a little closer using ACT (**exercise 4.6**).

Exercise 5.1/5.2

On the left is the table of putative autotransporters annotated in the *B. bronchiseptica* genome. You could now take a variety of routes to identify autotransporters in the other two genomes. We're going to start using a simple keyword search.

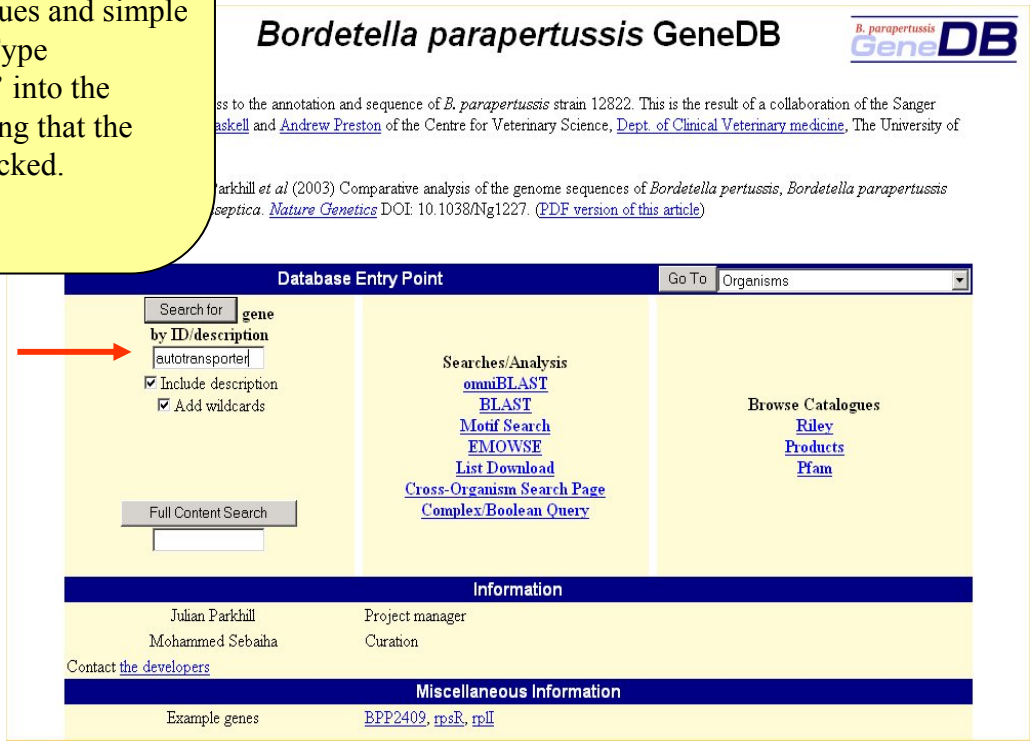
Table 2 Autotransporters encoded in the genomes of *B. pertussis*, *B. parapertussis* and *B. bronchiseptica*

	B. pertussis	*B. parapertussis*	*B. bronchiseptica*
SphB1			BB0419
Novel			BB0450
Novel			BB0452
Novel			BB0821
Novel			BB0916
Serum-resistance protein			BB0961
Pertactin			BB1366
BapA/AidB			BB1649
Vag8			BB1864
BapC			BB2033
Phg			BB2246
Novel			BB2270
SphB3 (serine protease)			BB2301
Novel			BB2324
SphB2			BB2741
Novel			BB2830
Novel			BB2941
Novel			BB3110
Novel			BB3111
TcfA			BB3291
BapB			BB3292

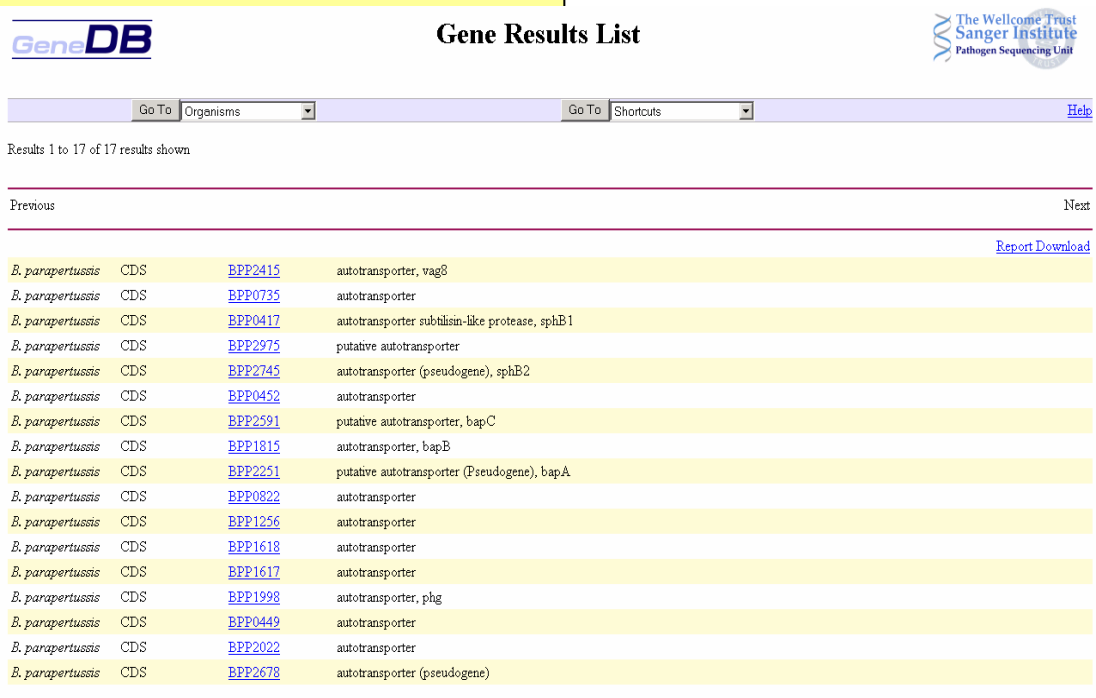
Pseudogenes are underlined.

Parkhill *et al.*, Nature Genetics (2003), 35: 32-40.

Start by going to the GeneDB homepage at www.genedb.org and choose to go to the *B. parapertussis* homepage by selecting this organism from the pull-down menu. This will get you to the species homepage, providing access to the data via tools (BLAST servers), browsable catalogues and simple search facilities. Type '**autotransporter**' into the search box, ensuring that the wild card box is ticked.



This returns a list of all genes with annotated product lines matching this search term.



GeneDB — Gene Results List — The Wellcome Trust Sanger Institute, Pathogen Sequencing Unit

Go To: Organisms | Go To: Shortcuts | Help

Results 1 to 17 of 17 results shown

Previous | Next

Report Download

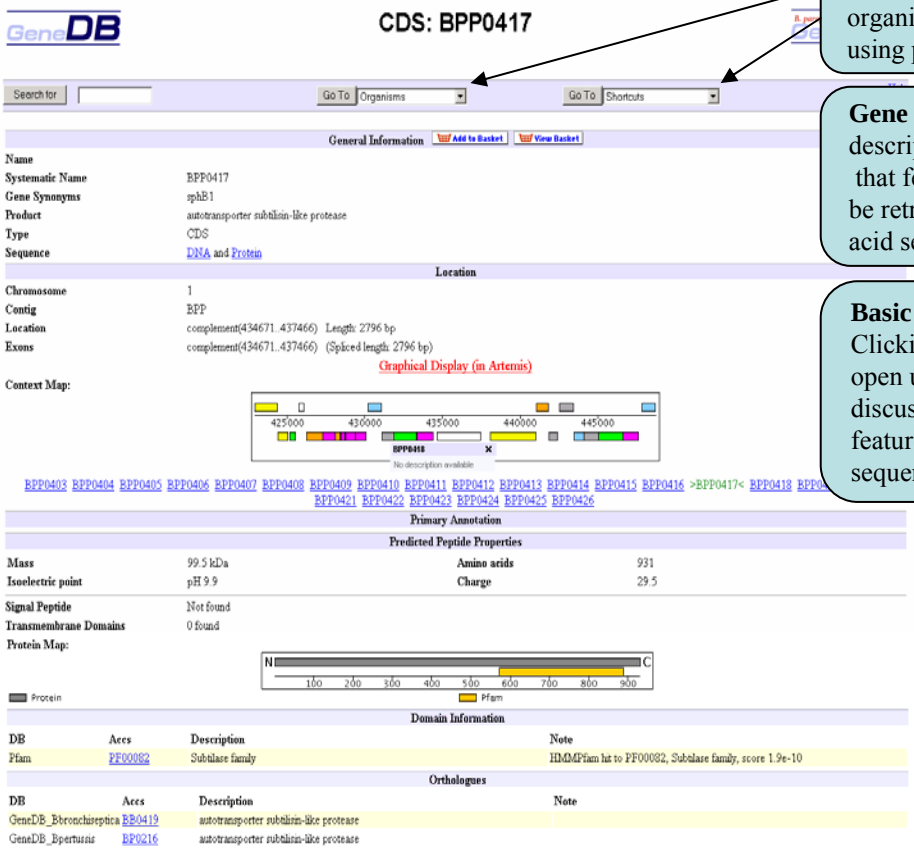
B. parapertussis	CDS	BPP2415	autotransporter, vag8
B. parapertussis	CDS	BPP0735	autotransporter
B. parapertussis	CDS	BPP0417	autotransporter subtilisin-like protease, sphB1
B. parapertussis	CDS	BPP2975	putative autotransporter
B. parapertussis	CDS	BPP2745	autotransporter (pseudogene), sphB2
B. parapertussis	CDS	BPP0452	autotransporter
B. parapertussis	CDS	BPP2591	putative autotransporter, bapC
B. parapertussis	CDS	BPP1815	autotransporter, bapB
B. parapertussis	CDS	BPP2251	putative autotransporter (Pseudogene), bapA
B. parapertussis	CDS	BPP0822	autotransporter
B. parapertussis	CDS	BPP1256	autotransporter
B. parapertussis	CDS	BPP1618	autotransporter
B. parapertussis	CDS	BPP1617	autotransporter
B. parapertussis	CDS	BPP1998	autotransporter, phg
B. parapertussis	CDS	BPP0449	autotransporter
B. parapertussis	CDS	BPP2022	autotransporter
B. parapertussis	CDS	BPP2678	autotransporter (pseudogene)

Click on '**BPP0417**' which will take you to the feature page, detailing information associated with this gene.

Navigation bar pull down menues: You can navigate between different organism datasets and search tools using pull down menus

Gene name and product information: The description lines are standardized and indexed so that features sharing the same description lines can be retrieved. Access to the nucleotide and amino acid sequences of the feature are also provided.

Basic location information and context map: Clicking on the 'Graphical display in Artemis' open up an Artemis applet – which will be discussed further in exercise 2. Via the applet, the feature can be viewed in the context of the sequence and additional annotation.



GeneDB — CDS: BPP0417

Search for | Go To: Organisms | Go To: Shortcuts

General Information

Name	
Systematic Name	BPP0417
Gene Synonyms	sphB1
Product	autotransporter subtilisin-like protease
Type	CDS
Sequence	DNA and Protein

Location

Chromosome	1
Contig	BPP
Location	complement(434671..437466) Length: 2796 bp
Exons	complement(434671..437466) (Spliced length: 2796 bp)

Graphical Display (in Artemis)

Context Map:

BPP0403 BPP0404 BPP0405 BPP0406 BPP0407 BPP0408 BPP0409 BPP0410 BPP0411 BPP0412 BPP0413 BPP0414 BPP0415 BPP0416 >BPP0417< BPP0418 BPP0421 BPP0422 BPP0423 BPP0424 BPP0425 BPP0426

Primary Annotation

Predicted Peptide Properties

Mass	99.5 kDa	Amino acids	931
Isoelectric point	pH 9.9	Charge	29.5
Signal Peptide	Not found		
Transmembrane Domains	0 found		

Protein Map:

Protein | Pfam

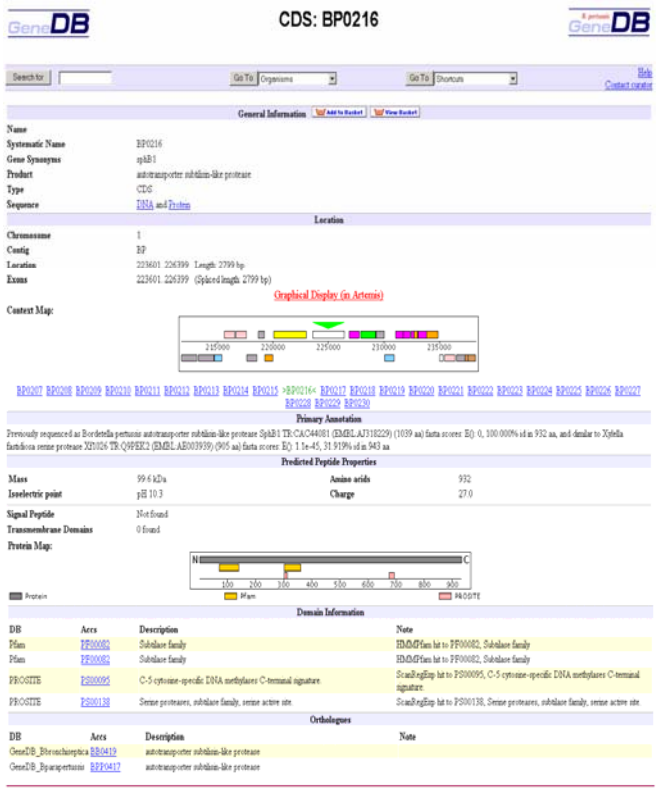
Domain Information

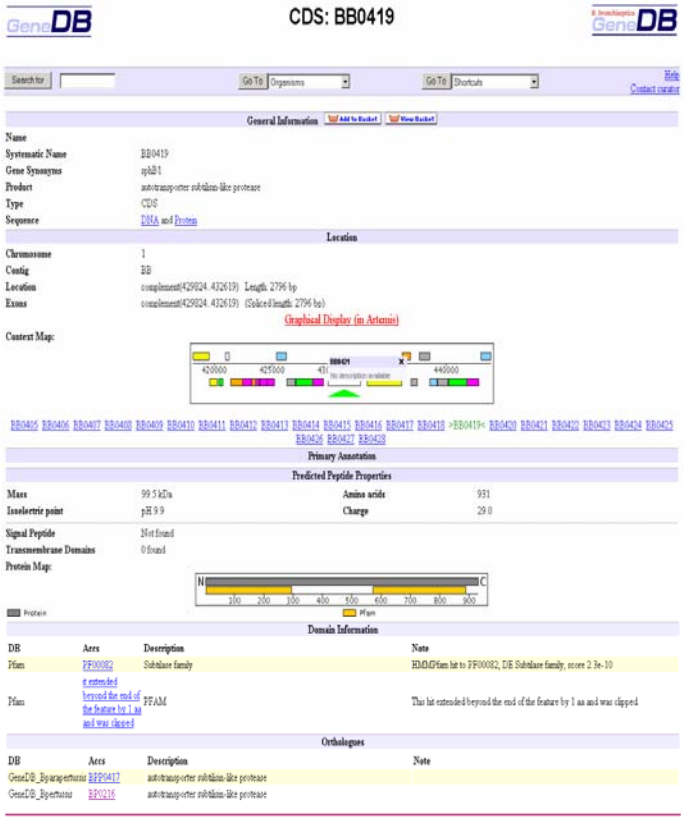
DB	Acc	Description	Note
Pfam	PF00082	Subtilase family	HMMPfam hit to PF00082, Subtilase family, score 1.9e-10

Orthologues

DB	Acc	Description	Note
GeneDB_Bbronchiseptica	BB0419	autotransporter subtilisin-like protease	
GeneDB_Bpertussis	BP0216	autotransporter subtilisin-like protease	

You'll find manually curated orthologue links towards the bottom of the feature page which will take you to the *B. pertussis* and *B. bronchiseptica* genes.



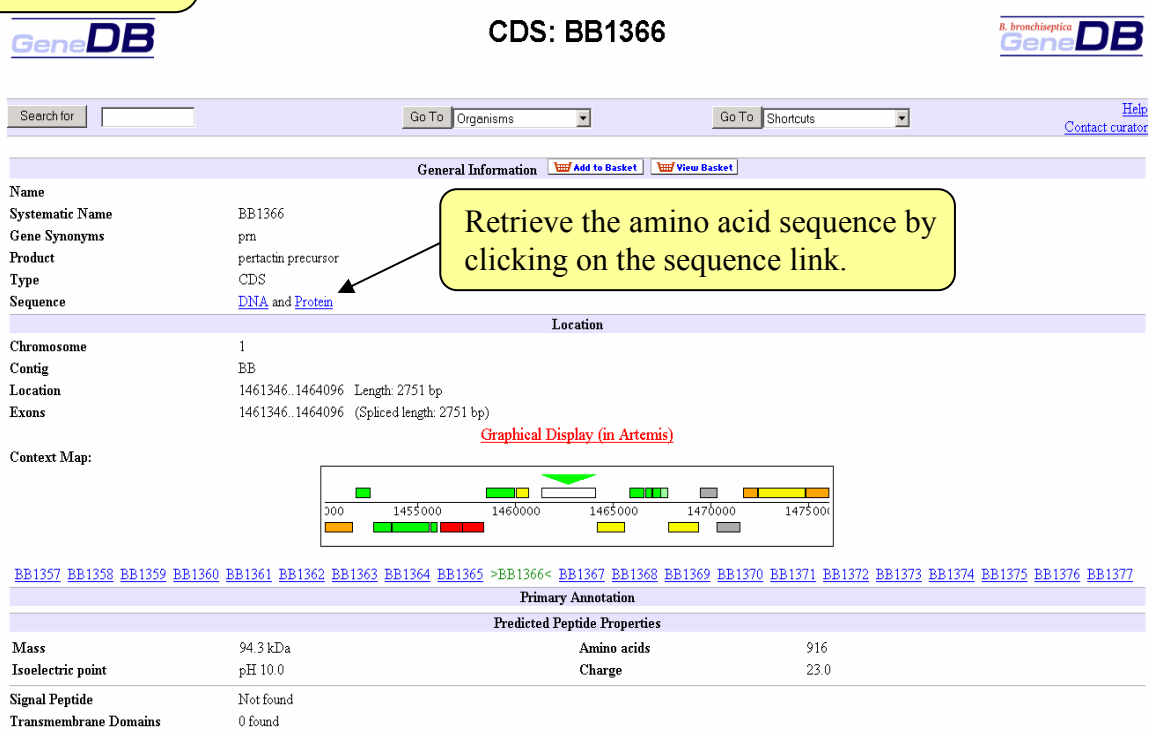


One way to identify the orthologues across the species would therefore be to access every *B. bronchiseptica* autotransporter and use the orthologue links. However, there are other means of identifying the putative orthologues.

Exercise 5.3

Using the BLAST server

Go to the *B. bronchiseptica* feature page for BB1366.



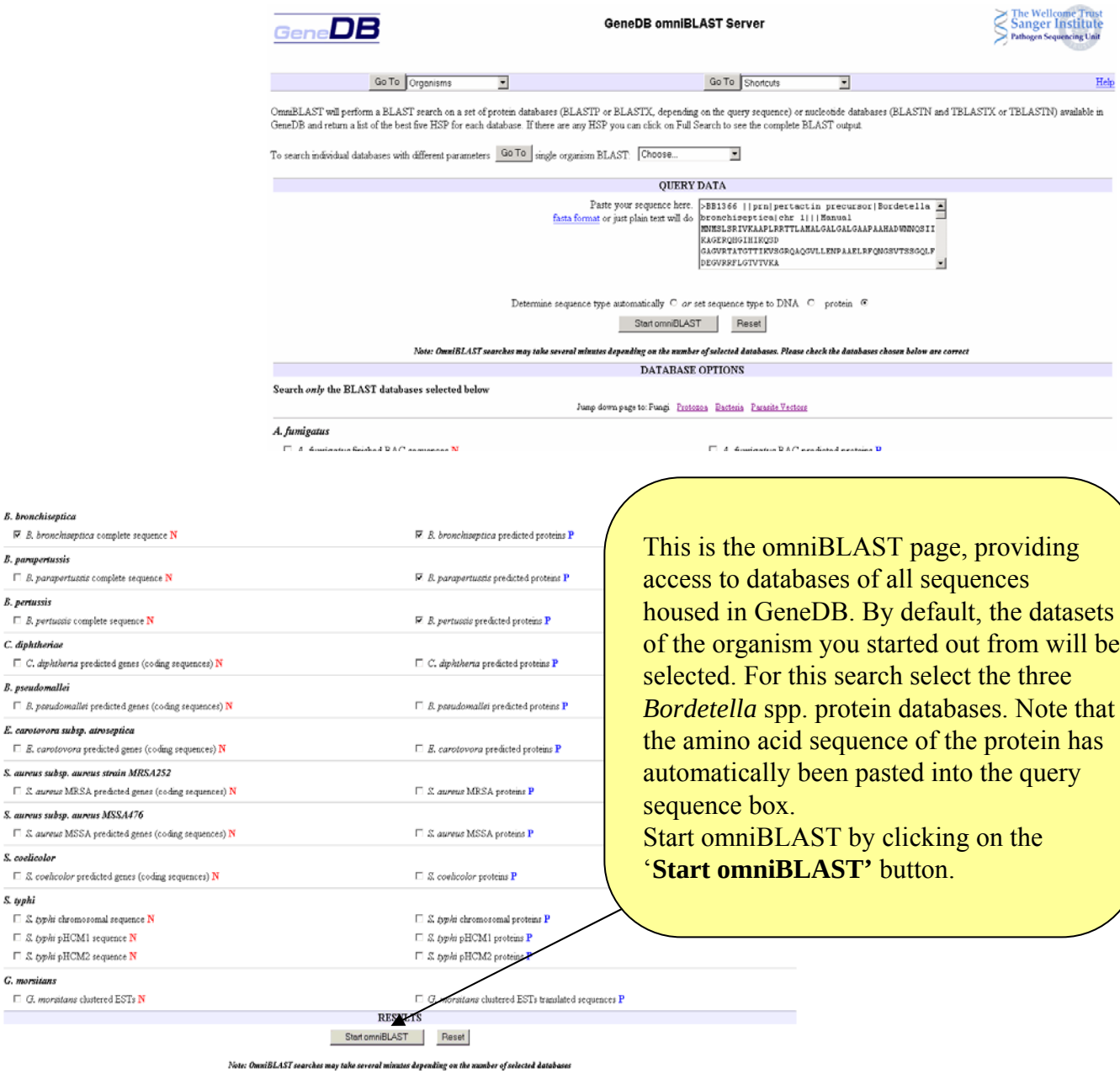
Retrieve the amino acid sequence by clicking on the sequence link.

Click on '**send to omniBLAST**' link. omniBLAST permits searching across different databases selected by the user.

```
GCCAACACCA TGCTGCTGGT GCAGACGCCA CGAGGCAGCG CGGCGACCTT TACCCTTGCC
AACAAGGACG GCAAGGTCGA TATCGGTACC TACCGCTATC GATTGGCCGC CAACGGCAAT
GGGCAGTGGA GCCTGGTGGG CGCGAAGGCG CCGCCGGCGC CCAAGCCCGC GCCGCAGCCC
GGTCCCCAGC CCGGTCCCCA GCCGCCGCAG CCGCCGCAGC CGCCACAGAG GCAGCCGGAA
GCGCCGGCGC CGCAACCGCC GGCGGGCAGG GAGTTGTCCG CCGCCGCCAA CGCGGCGGTC
AACACGGGTG GGGTGGGCCT GGCCAGCACG CTCTGGTACG CCGAAAGCAA TGCGTTGTCC
AAGCGCCTGG GCGAGTTGCG CCTGAATCCG GACGCCGGCG GCGCTTGGGG CCGCGGCTTC
GCGCAACGCC AGCAACTGGA CAACCGCGCC GGGCGGCGCT TCGACCAGAA GGTGGCCGGC
TTCGAGCTGG GCGCCGACCA CGCGGTGGCG GTGGCCGGCG GGCGCTGGCA CCTGGGCGGG
CTGGCCGGCT ATACGCGCGG CGACCGCGGC TTTACCGGCG ACGGCGGCGG CCACACCGAC
AGCGTGCATG TCGGGGGCTA TGCCACCTAT ATCGCCAACA GCGGTTTCTA CCTGGACGCG
ACGCTGCGCG CCAGCCGCCT CGAAAATGAC TTCAAGGTGG CGGGCAGCGA TGGGTACGCG
GTCAAGGGCA AGTACCGCAC CCATGGGGTA GGCGCCTCGC TCGAGGCGGG CCGGCGCTTC
GCCCATGCCG ACGGCTGGTT CCTCGAGCCG CAGGCCGAGC TGGCGGTGTT CCGGGTCGGC
GGCGGTGCGT ACCGCGCGGC CAATGGCCTG CGGGTGCGCG ACGAAGGCGG CAGCTCGGTG
CTGGGTCGCC TGGGCCTGGA GGTCGGCAAG CGCATCGAAC TGGCAGGCGG CAGGCAGGTG
CAGCCATACA TCAAGGCCAG CGTGTTGCAG GAGTTCGACG GCGCGGGTAC GGTACGCACC
AACGGCATCG CGCACCGCAC CGAACTGCGC GGCACGCGCG CCGAACTGGG CCTGGGCATG
GCCGCCGCGC TGGGCCGCGG CCACAGCCTG TATGCCTCGT ACGAGTACTC CAAGGGCCCG
AAGCTGGCCA TGCCGTGGAC CTTCCACGCG GGCTACCGGT ACAGCTGGTA A
```

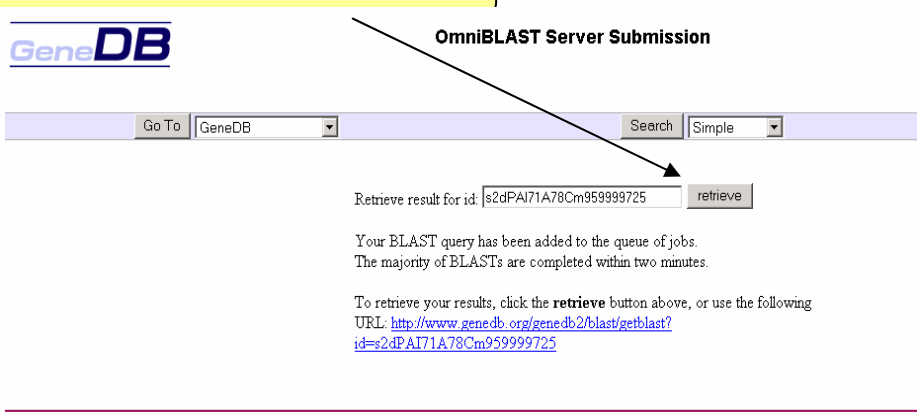
Send to GeneDB omniBLAST Send to GeneDB BLAST Send to BLAST at NCBI

```
>BB1366 ||prn|pertactin precursor|Bordetella bronchiseptica|chr 1|||Manual
MNMSLSRIVK AAPLRRTTLA MALGALGALG AAPAAHADWN NQSIIKAGER QHGIHIKQSD
GAGVRTATGT TIKVSGRQAQ GVLLENPAAE LRFQNGSVTS SGQLFDEGVR RFLGTVTVKA
GKLVADHATL ANVSDTRDDD GIALYVAGEQ AQASIADSTL QGAGGVRVER GANVTVQRST
IVDGGLHIGT LQPLQPEDLP PSRVVLGDTS VTAVPASGAP AAVSVFGANE LTVDGGHITG
GRAAGVAAMD GAIVHLQRAT IRRGDAPAGG AVPGGAVPGG AVPGGFGPLL DGWYGVDVSD
STVDLAQSIV EAPQLGAAIR AGRGARVTVS GGSLSAPHGN VIETGGGARR FPPPASPLSI
TLQAGARAQG RALLYRVLPE PVKLTLAGGA QGQGDIVATE LPPIPGASSG PLDVALASQA
RWTGATRAVD SLSIDNATWV MTDNSNVGAL RLASDGSVDF QQPAEAGRFK VLMVDTLAGS
GLFRMNVFAD LGLSDKLVVM RDASGQHRLW VRNSGSEPAS ANTMLLVQTP RGSAATFTLA
NKDGKVDIGT YRYRLAANGN GQWSLVGAKA PPAPKPAPQP GPQPGPQPPQ PPQPPQRQPE
APAPQPPAGR ELSAAANAAV NTGGVGLAST LWYAESNALS KRLGELRLNP DAGGAWGRGF
AQRQQLDNRA GRRFDQKVAG FELGADHAVA VAGGRWHLGG LAGYTRGDRG FTGDGGGHTD
SVHVGGYATY IANSGFYLDA TLRASRLEND FKVAGSDGYA VKGKYRTHGV GASLEAGRRF
AHADGWFLEP QAELAVFRVG GGAYRAANGL RVRDEGGSSV LGRLGLEVGK RIELAGGRQV
QPYIKASVLQ EFDGAGTVRT NGIAHRTELR GTRAELGLGM AAALGRGHSL YASYEYSKGP
KLAMPWTFHA GYRYSW
```

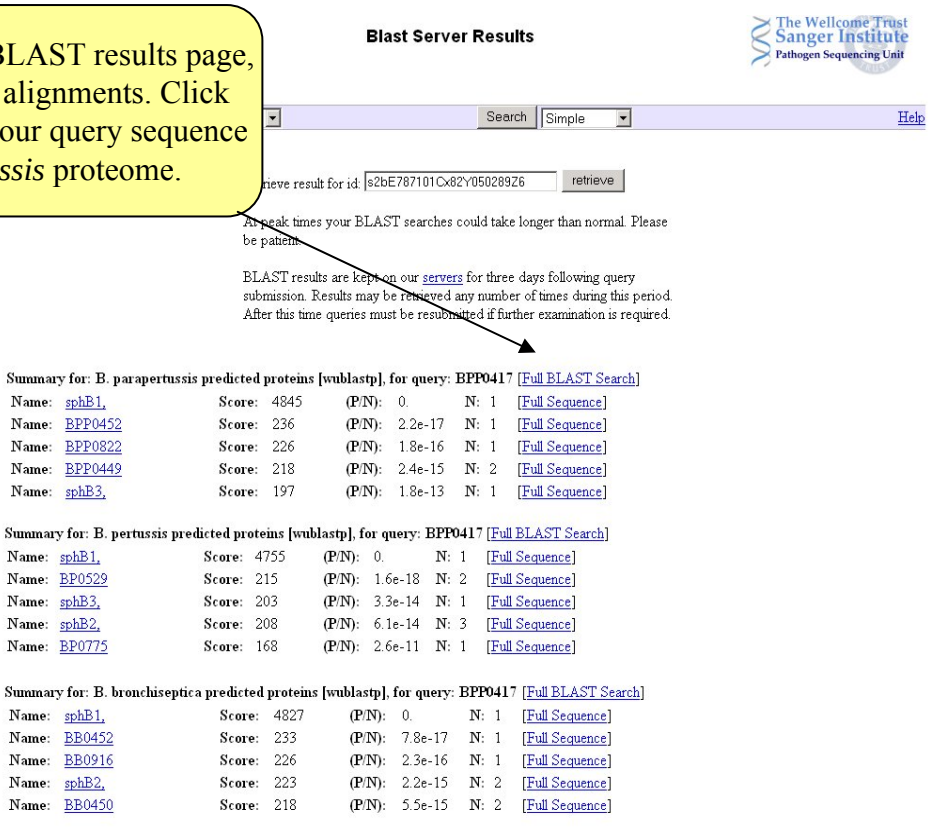


This is the omniBLAST page, providing access to databases of all sequences housed in GeneDB. By default, the datasets of the organism you started out from will be selected. For this search select the three *Bordetella* spp. protein databases. Note that the amino acid sequence of the protein has automatically been pasted into the query sequence box.
Start omniBLAST by clicking on the '**Start omniBLAST'** button.

Retrieve the results by clicking on the '**retrieve**' button. As indicated, results will be accessible for the next 2 weeks using the listed URL.



You will retrieve an abbreviated BLAST results page, listing only the top 5 hits without alignments. Click on the '**Full BLAST Search**' of your query sequence against the predicted *B. parapertussis* proteome.



Blast Server Results

The Wellcome Trust Sanger Institute
Pathogen Sequencing Unit

Search | Simple | Help

ieve result for id: s2bE787101Cx82Y050289Z6 retrieve

At peak times your BLAST searches could take longer than normal. Please be patient.

BLAST results are kept on our servers for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

Summary for: B. parapertussis predicted proteins [wublastp], for query: BPP0417 [Full BLAST Search]

Name:	sphB1,	Score:	4845	(P/N):	0.	N: 1	[Full Sequence]
Name:	BPP0452	Score:	236	(P/N):	2.2e-17	N: 1	[Full Sequence]
Name:	BPP0822	Score:	226	(P/N):	1.8e-16	N: 1	[Full Sequence]
Name:	BPP0449	Score:	218	(P/N):	2.4e-15	N: 2	[Full Sequence]
Name:	sphB3,	Score:	197	(P/N):	1.8e-13	N: 1	[Full Sequence]

Summary for: B. pertussis predicted proteins [wublastp], for query: BPP0417 [Full BLAST Search]

Name:	sphB1,	Score:	4755	(P/N):	0.	N: 1	[Full Sequence]
Name:	BP0529	Score:	215	(P/N):	1.6e-18	N: 2	[Full Sequence]
Name:	sphB3,	Score:	203	(P/N):	3.3e-14	N: 1	[Full Sequence]
Name:	sphB2,	Score:	208	(P/N):	6.1e-14	N: 3	[Full Sequence]
Name:	BP0775	Score:	168	(P/N):	2.6e-11	N: 1	[Full Sequence]

Summary for: B. bronchiseptica predicted proteins [wublastp], for query: BPP0417 [Full BLAST Search]

Name:	sphB1,	Score:	4827	(P/N):	0.	N: 1	[Full Sequence]
Name:	BB0452	Score:	233	(P/N):	7.8e-17	N: 1	[Full Sequence]
Name:	BB0916	Score:	226	(P/N):	2.3e-16	N: 1	[Full Sequence]
Name:	sphB2,	Score:	223	(P/N):	2.2e-15	N: 2	[Full Sequence]
Name:	BB0450	Score:	218	(P/N):	5.5e-15	N: 2	[Full Sequence]

GeneDB

Blast Server Results

The Wellcome Trust Sanger Institute
Pathogen Sequencing Unit

Go To | GeneDB | Search | Simple | Help

Retrieve result for id: s2bA7796c065Cg7239rY retrieve

At peak times your BLAST searches could take longer than normal. Please be patient.

BLAST results are kept on our servers for three days following query submission. Results may be retrieved any number of times during this period. After this time queries must be resubmitted if further examination is required.

Low complexity filtering disabled
Repeatmasker disabled

BLASTP 2.0MP-WashU [16-Sep-2002] [decunix4.0-ev6-I32LPF64 2002-09-18T19:28:12]

Copyright (C) 1996-2002 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

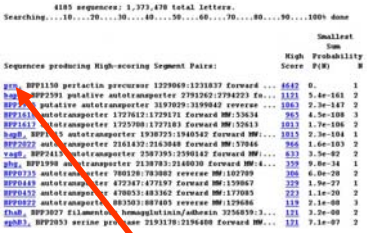
Reference: Gish, W. (1996-2002) http://blast.wustl.edu

Query= BB1366
(916 letters)
Database: BPa.pep
4185 sequences; 1,373,478 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
prn, BPP1150 pertactin precursor 1229069:1231837 forward ...	4642	0.	1
bapC, BPP2591 putative autotransporter 2791262:2794223 fo...	1121	5.4e-161	2
BPP2975 putative autotransporter 3197029:3199842 reverse ...	1063	2.3e-147	2
BPP1618 autotransporter 1727612:1729171 forward MW:53634	965	4.5e-108	3
BPP1617 autotransporter 1725708:1727183 forward MW:52613	1013	1.7e-106	2
bapB, BPP1815 autotransporter 1938725:1940542 forward MW:...	1015	2.3e-104	1
BPP2022 autotransporter 2161432:2163048 forward MW:57046	966	1.6e-103	2
vagB, BPP2415 autotransporter 2587395:2590142 forward MW:...	633	3.5e-82	2
phg, BPP1990 autotransporter 2130703:2140030 forward MW:4...	559	9.0e-34	1
BPP0735 autotransporter 780128:783082 reverse MW:102709	306	6.0e-28	2
BPP0449 autotransporter 472347:477197 forward MW:159867	329	1.9e-27	1
BPP0452 autotransporter 478053:483362 forward MW:177085	223	1.1e-20	2
BPP0822 autotransporter 883503:887405 reverse MW:129686	119	2.1e-08	3
fhaB, BPP3027 filamentous hemagglutinin/adhesin 3256839:3...	121	3.2e-08	2
sphB3, BPP2053 serine protease 2193178:2196408 forward MW:...	121	7.1e-07	2
BPP3791 Proline-rich inner membrane protein 4104773:41058...	126	6.7e-06	2
BPP2360 putative membrane protein 2530013:2532498 reverse...	116	8.1e-06	2
fhaI, BPP2489 adhesin 2668735:2681391 reverse MW:439845	111	8.2e-06	2
fhaS, BPP1243 adhesin 1331868:1339673 forward MW:268245	110	0.00017	3
BPP0732 hypothetical protein 776068:777270 forward MW:40873	113	0.00017	2
BPP1520 conserved hypothetical protein 1626284:1627114 fo...	103	0.00032	1
tonB, BPP2535 siderophore-mediated iron transport protein...	102	0.00042	1
ftsY, BPP0849 cell division protein 918936:920003 forward...	103	0.00051	1
flik, flaE, flaR, BPP1505 flagellar hook-length control p...	80	0.00082	2
dnaX, dnaZ, dnazX, BPP1221 DNA polymerase III subunit tau...	105	0.0046	3
BPP3783 hypothetical protein 4096537:4098510 forward MW:6...	72	0.0072	2
BPP2382 putative exported protein 2545322:2547705 forward...	83	0.0077	1

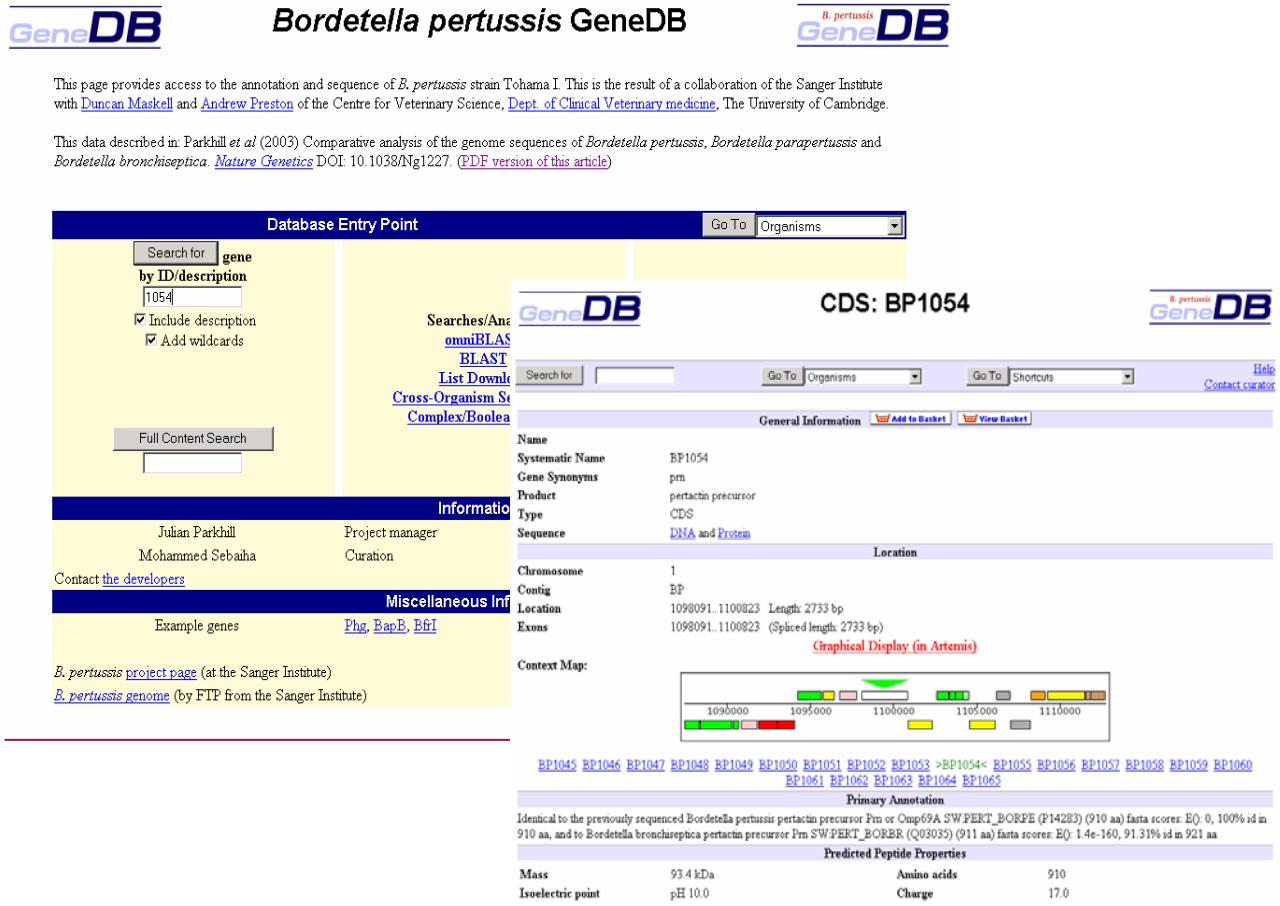
The BLAST result reveals 13 genes In *B. parapertussis* which have a High degree of sequence similarity. Notice they are all annotated as Autotransporters.

```
>prn, BPP1150 pertactin precursor 1229069:1231837 forward MW:95179
[Full Sequence] 2

        Length = 922

 Score = 4642 (1639.1 bits), Expect = 0., P = 0.
 Identities = 909/925 (98%), Positives = 910/925 (98%)
[HSP Sequence]

Query:     1 MNMSLSRIVKAAPLRRTTLAMALGALGALGAAPAAHADWNNQSIIKAGERQHGIHIKQSD 60
             MNMSLSRIVKAAPLRRTTLAMALGALGA   APAA+ADWNNQSIIKAGERQHGIHIKQSD
Sbjct:     1 MNMSLSRIVKAAPLRRTTLAMALGALGA---APAAYADWNNQSIIKAGERQHGIHIKQSD 57

Query:    61 GAGVRTATGTTIKVSGRQAQGVLLENPAAELRFQNGSVTSSGQLFDEGVRRFLGTVTVKA 120
             GAGVRTATGTTIKVSGRQAQGVLLENPAAELRFQNGSVTSSGQLFDEGVRRFLGTVTVKA
Sbjct:    58 GAGVRTATGTTIKVSGRQAQGVLLENPAAELRFQNGSVTSSGQLFDEGVRRFLGTVTVKA 117

Query:   121 GKLVADHATLANVSDTRDDDGIALYVAGEQAQASIADSTLQGAGGVRVERGANVTVQRST 180
             GKLVADHATLANVSDTRDDDGIALYVAGEQAQASIADSTLQGAGGVRVERGANVTVQRST
Sbjct:   118 GKLVADHATLANVSDTRDDDGIALYVAGEQAQASIADSTLQGAGGVRVERGANVTVQRST 177

Query:   181 IVDGGLHIGTLQPLQPEDLPPSRVVLGDTSVTAVPASGAPAAVSVFGANELTVDGGHITG 240
             IVDGGLHIGTLQPLQPEDLPPSRVVLGDTSVTAVPASGAPAAV VFGANELTVDGGHITG
Sbjct:   178 IVDGGLHIGTLQPLQPEDLPPSRVVLGDTSVTAVPASGAPAAVFVFGANELTVDGGHITG 237

Query:   241 GRAAGVAAMDGAIVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPLLDGWYGVDVSD 300
             GRAAGVAAMDGAIVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPLLDGWYGVDVSD
Sbjct:   238 GRAAGVAAMDGAIVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPLLDGWYGVDVSD 297

Query:   301 STVDLAQSIVEAPQLGAAIRAGRGARVTVSGGSLSAPHGNVIETGGGARRFPPPASPLSI 360
             STVDLAQSIVEAPQLGAAIRAGRGARVTVSGGSLSAPHGNVIETGGGARRFPPPASPLSI
Sbjct:   298 STVDLAQSIVEAPQLGAAIRAGRGARVTVSGGSLSAPHGNVIETGGGARRFPPPASPLSI 357

Query:   361 TLQAGARAQGRALLYRVLPEPVKLTLAGGAQGQGDIVATELPPIPGASSGPLDVALASQA 420
             TLQAGARAQGRALLYRVLPEPVKLTLAGGAQGQGDIVATELPPIPGASSGPLDVALASQA
Sbjct:   358 TLQAGARAQGRALLYRVLPEPVKLTLAGGAQGQGDIVATELPPIPGASSGPLDVALASQA 417

Query:   421 RWTGATRAVDSLSIDNATWVMTDNSNVGALRLASDGSVDFQQPAEAGRFKVLMVDTLAGS 480
             RWTGATRAVDSLSIDNATWVMTDNSNVGALRLASDGSVDFQQPAEAGRFKVLMVDTLAGS
Sbjct:   418 RWTGATRAVDSLSIDNATWVMTDNSNVGALRLASDGSVDFQQPAEAGRFKVLMVDTLAGS 477
```

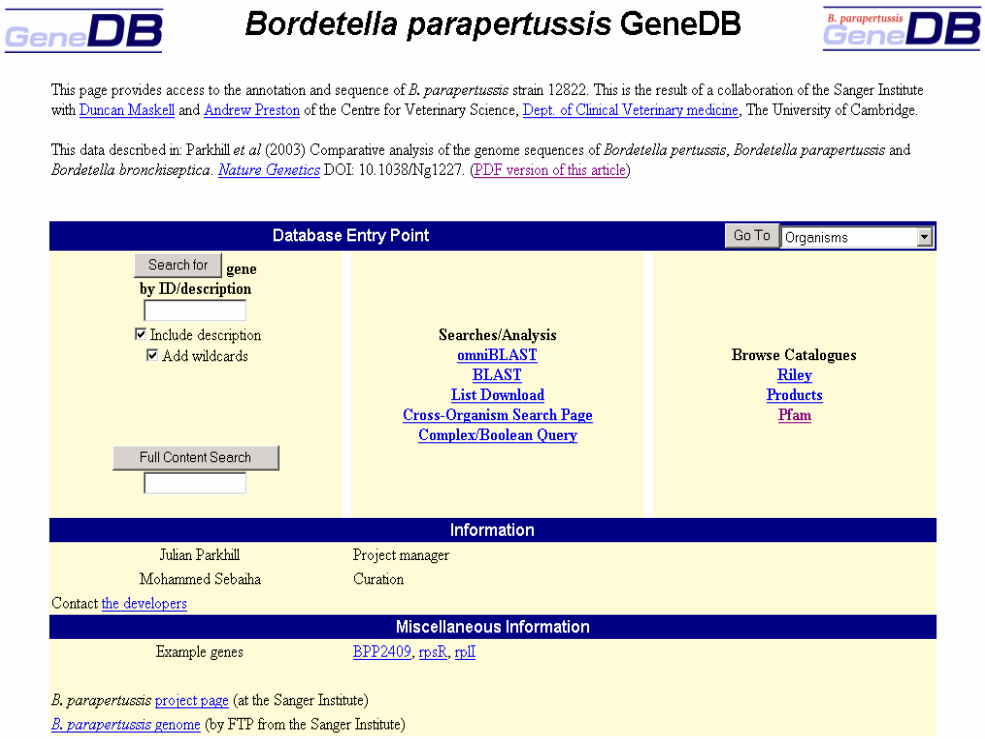


Selecting the link for the top BLAST hit will take you to the Multiple alignment of the *B. bronchiseptica* query sequence and the likely ortholog in *B. parapertussis.*

Open a new window and go to the *B. pertussis* homepage and you can access the putative orthologue by typing in the gene name.

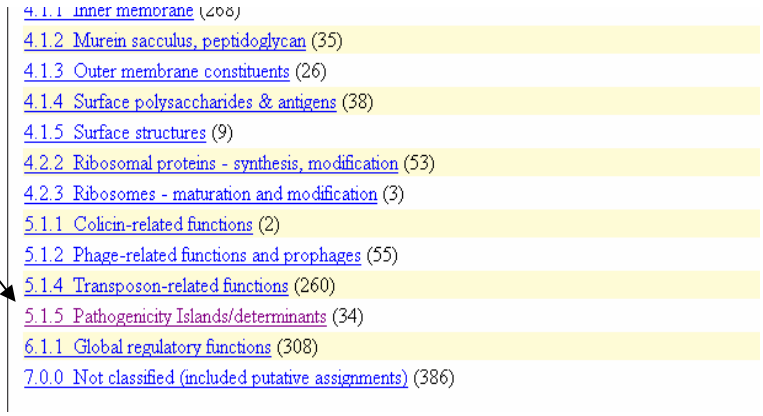


Exercise 5.4 Using browsable catalogues (Riley)



GeneDB provides access to browsable catalogues (product lines, Pfam predictions, Riley classification). Click on 'Riley' which will take you to a list of each of the categories linked to genes annotated to this term.

Click on 'Pathogenicity Islands/determinants'



This brings up a list of all the genes annotated to this term. Are the autotransporters you're looking for listed here?

Report Download

B. pertussis	CDS	BP0216	autotransporter subtilisin-like protease, sphB1
B. pertussis	CDS	BP0529	autotransporter
B. pertussis	CDS	BP0758	cyclolysin-activating lysine-acyltransferase, cyaC, HlyC
B. pertussis	CDS	BP0760	bifunctional hemolysin-adenylate cyclase precursor, cyaA, cya
B. pertussis	CDS	BP0761	cyclolysin secretion ATP-binding protein, cyaB
B. pertussis	CDS	BP0762	cyclolysin secretion protein, cyaD
B. pertussis	CDS	BP0763	cyclolysin secretion protein, cyaE
B. pertussis	CDS	BP0874	vir-repressed protein, vir-18
B. pertussis	CDS	BP1054	pertactin precursor, prn
B. pertussis	CDS	BP1110	serine protease, sphB3
B. pertussis	CDS	BP1112	putative outer membrane ligand binding protein, bipA
B. pertussis	CDS	BP1200	autotransporter (pseudogene), bapB
B. pertussis	CDS	BP1201	tracheal colonization factor precursor, tcfA
B. pertussis	CDS	BP1251	putative toxin
B. pertussis	CDS	BP1344	autotransporter
B. pertussis	CDS	BP1610	putative autotransporter (Pseudogene)
B. pertussis	CDS	BP1660	autotransporter, sphB2
B. pertussis	CDS	BP1767	autotransporter, phg
B. pertussis	CDS	BP1793	autotransporter (Pseudogene)
B. pertussis	CDS	BP1884	hemolysin activator-like protein, fhaC
B. pertussis	CDS	BP2224	putative autotransporter, bapA
B. pertussis	CDS	BP2315	autotransporter, vag8
B. pertussis	CDS	BP2468	Virulence protein, vrg-6
B. pertussis	CDS	BP2627	autotransporter (pseudogene)
B. pertussis	CDS	BP2667	adhesin, fhaS
B. pertussis	CDS	BP2738	autotransporter (pseudogene), bapC

Out of interest, have a look at the Pfam/product browsable catalogues. The Pfam domain of interest is '**Autotransporter beta-domain (PF03797)**'. Can you identify autotransporters across the 3 genomes that way?

Note that you can also download this list inc. the sequences by clicking on the '**Report Download**' button.

GeneDB Id List Information

Sanger Institute Pathogen Sequencing Unit

Go To Organisms | Go To Shortcuts | Help

ID List

Please enter your gene/rna ids either:

- as database cross-references (eg in the form GeneDB_Spombe:SPAC1002.09c).
- or you can just use the systematic ids eg (Tb927.1.710) **but** in this case you must also set the default organism.

Default Organism: B. pertussis

```
GeneDB_Bpertussis:BP0216
GeneDB_Bpertussis:BP0529
GeneDB_Bpertussis:BP0758
GeneDB_Bpertussis:BP0760
GeneDB_Bpertussis:BP0761
GeneDB_Bpertussis:BP0762
GeneDB_Bpertussis:BP0763
GeneDB_Bpertussis:BP0874
GeneDB_Bpertussis:BP1054
GeneDB_Bpertussis:BP1110
GeneDB_Bpertussis:BP1112
GeneDB_Bpertussis:BP1200
GeneDB_Bpertussis:BP1201
GeneDB_Bpertussis:BP1251
GeneDB_Bpertussis:BP1344
GeneDB_Bpertussis:BP1610
GeneDB_Bpertussis:BP1660
GeneDB_Bpertussis:BP1767
GeneDB_Bpertussis:BP1793
GeneDB_Bpertussis:BP1884
```

Information Required For Each RNA/CDS

- (•) Descriptions
- ( ) DNA (Unspliced sequence) or cDNA
- ( ) DNA (Spliced sequence)
- ( ) Protein sequence
- ( ) Intergenic Sequence (3')
- ( ) Intergenic Sequence (5')

Number of bases: 20

- ( ) Sequence Range

3' distance: 0 5' distance: 0

- ( ) Orthologues

Submit Query | Reset

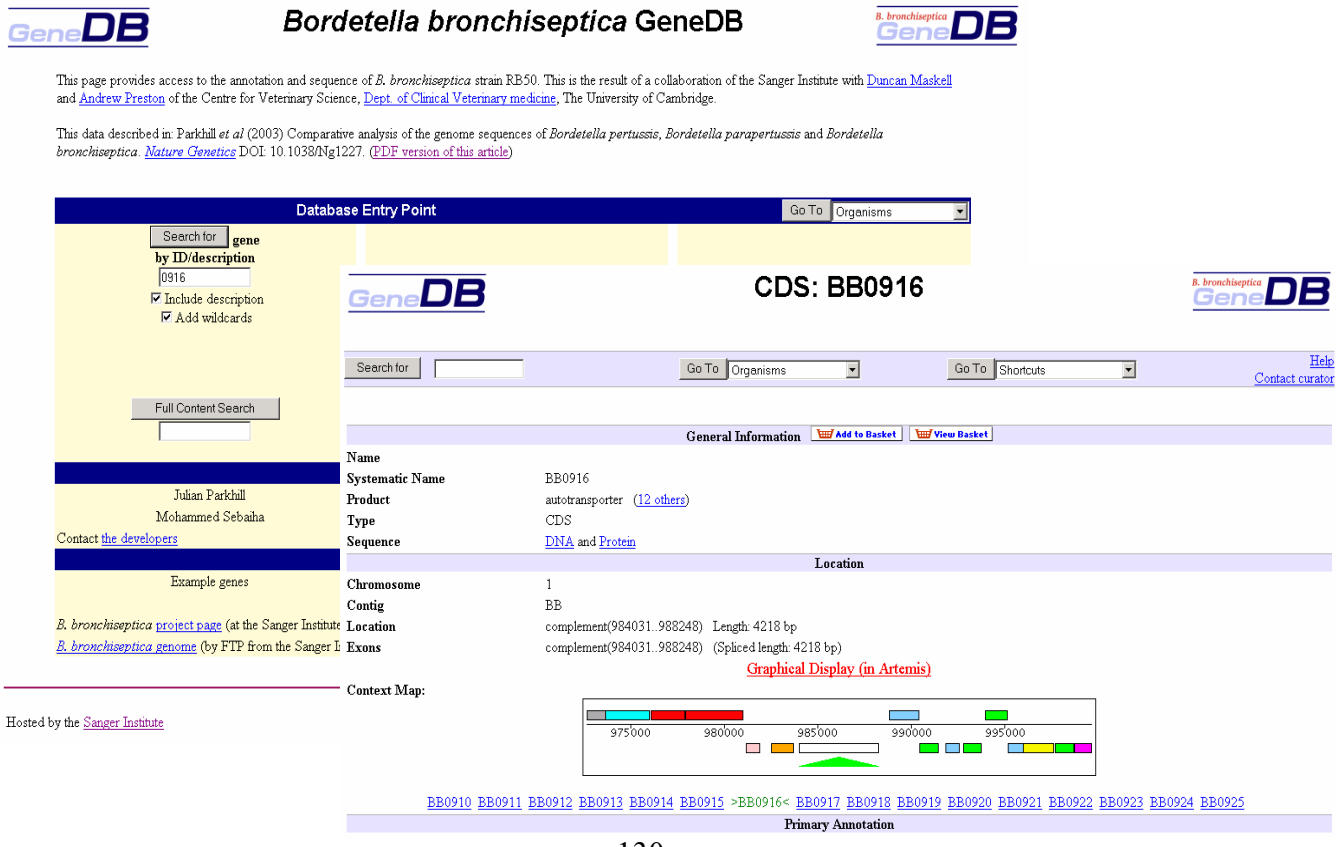
This is what your completed list should look like.

Table 2 Autotransporters encoded in the genomes of *B. pertussis*, *B. parapertussis* and *B. bronchiseptica*

	B. pertussis	*B. parapertussis*	*B. bronchiseptica*
SphB1	BP0216	BPP0417	BB0419
Novel	-	BPP0449	BB0450
Novel	BP0529	BPP0452	BB0452
Novel	-	BPP0735	BB0821
Novel	-	BPP0822	BB0916
Serum-resistance protein	BP3494	BPP0867	BB0961
Pertactin	BP1054	BPP1150	BB1366
BapA/AidB	BP2224	BPP2251	BB1649
Vag8	BP2315	BPP2415	BB1864
BapC	BP2738	BPP2591	BB2033
Phg	BP1767	BPP1998	BB2246
Novel	BP1793	BPP2022	BB2270
SphB3 (serine protease)	BP1110	BPP2053	BB2301
Novel	BP2627	BPP1256	BB2324
SphB2	BP1660	BPP2745	BB2741
Novel	BP1344	BPP2678	BB2830
Novel	BP1610	BPP2975	BB2941
Novel	–	BPP1618	BB3110
Novel	–	BPP1617	BB3111
TcfA	BP1201	–	BB3291
BapB	BP1200	BPP1815	BB3292

Pseudogenes are underlined.

Exercise 5.5 Now imagine you're particularly interested in the genomic loci around BB0916 as it only appears to have an orthologue in *B. parapertussis* and not *B. pertussis*. As briefly mentioned earlier, GeneDB supports an Artemis applet with which you can view sequence in more detail as well as being able to download sequence.



Identifying monosaccharide transporters within the genomes of *P. falciparum*, *P. Berghei* and *P. chabaudi*

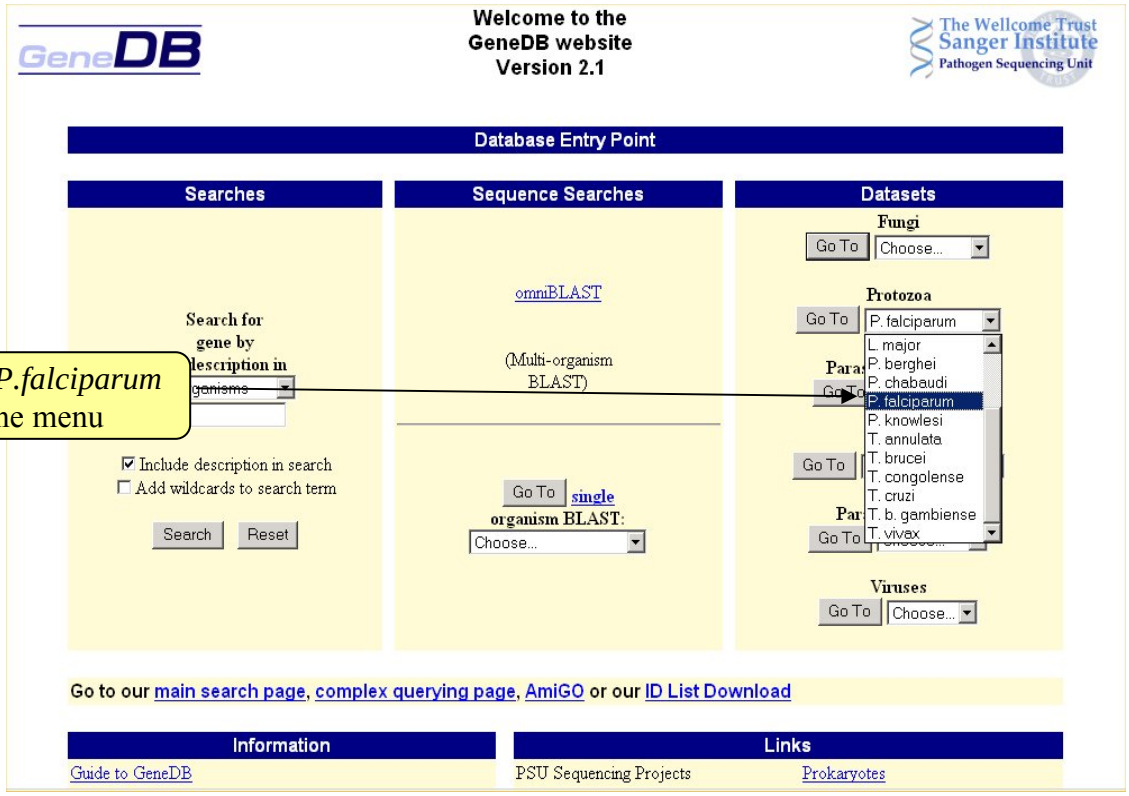
1

The following exercises aim to introduce you to the features that allow quick and convenient data mining from GeneDB, and will equip you with the tools to use the database to facilitate your own research. Hopefully, they will also make you aware of its strengths and limitations, and highlight the advantage of using several search strategies.
The aim is to identify monosaccharide transporters in *Plasmodium falciparum*, *Plasmodium berghei* and *Plasmodium chabaudi*..Glucose transporters are promising drug targets as asexual-stage parasites depend heavily upon glucose for energy (Joet *et. al.*). Compounds such as O-methylderivatives of glucose, have been shown to selectively inhibit glucose transport by *Plasmodium falciparum.* Hexose Transporter (PfHT). Using a variety of tools/methods, some of which you will already have covered in earlier modules, you'll identify this gene in *Plasmodium falciparum* and then go on to identify its putative orthologues in *P. berghei* and *P. chabaudi*. This would obviously be of interest to a researcher in the field who wanted to assess how similar the putative homologues were to the gene in *Plasmodium falciparum*.
The following are key references that will be provided to you to give you some background information. They are only for your reference and for the purpose of this exercise reading the abstract is probably sufficient.

Joet *et al*., Comparative characterisation of hexose transporters of *Plasmodium knowlesi*, *Plasmodium yoelli* and *Toxoplasma gondii* highlights functional differences within the apicomplexan family. PMID12238947. Biochem J. 2002 (Dec) pp 923-9.
Krishna *et al*., Transport processes in *P. falciparum*-infected erythrocytes: potential as new drug targets. PMID:12435441. Int J Parisitol (Dec) pp 1567-73.

Exercise 1 Searching GeneDB using simple keyword searches

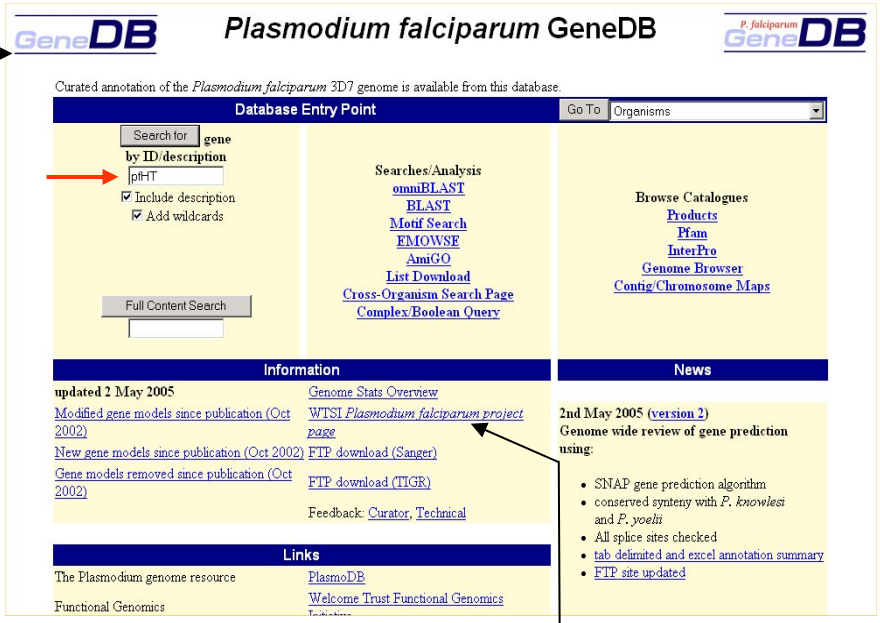
2 Go to the GeneDB homepage (http://www.genedb.org)



3 Select *P.falciparum* from the menu

4 Try **pfHT** as a search term by typing into the search box (red arrow) then pressing return or clicking the search for button. Ensure that the '**Add wildcards to search term**' box is also selected. This will search all the gene names and product/description lines assigned to features within the *P. falciparum* dataset. Not selecting the wildcard box would only return exact matches. After examining the gene page for pfHT (on page 5), use the browser back button to return to the search page and try the following search terms sugar transporter, monosacharide transporter, hexose transporter and transporter.

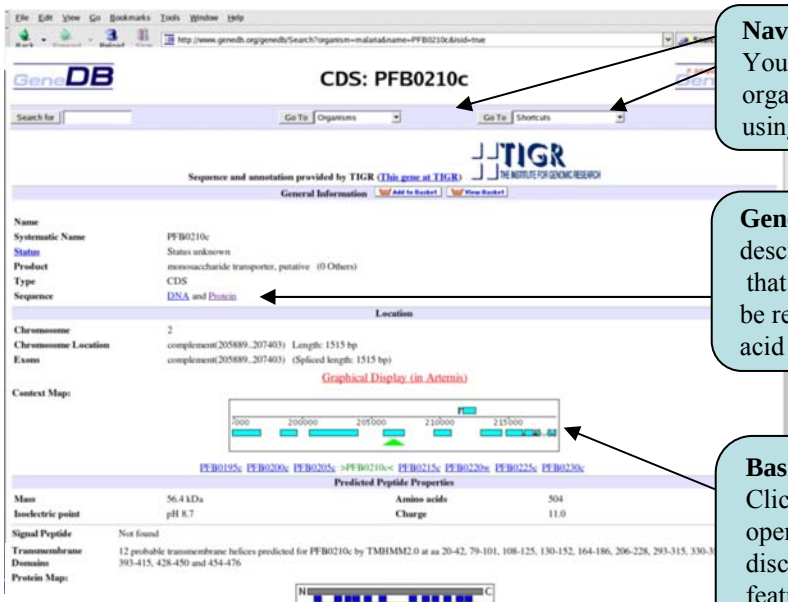
Note: clicking on the GeneDB icon will take you back to the GeneDB home page



Clicking on the link – *Status and Project Information* - will take you to pages that describe the status and background of protozoan genome projects and other genome projects at the WTSI. Have a look at the *Plasmodium* projects for reference.

5 The results for a search using '**pfHT**' should take you to the gene page for **PFB0210c (next page),** while the other searches if there are multiple results will Display a results list.

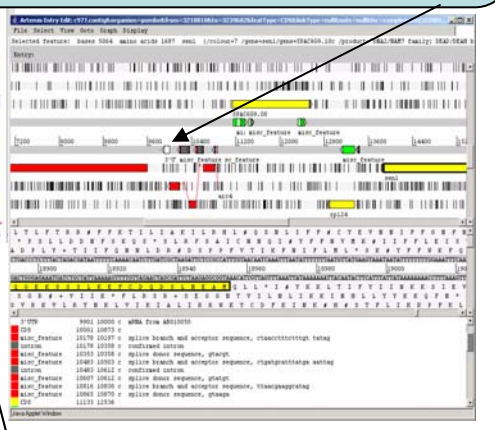
6 Click on the links on the feature page to see how the data are cross-linked and referenced.



Navigation bar pull down menues: You can navigate between different organism datasets and search tools using pull down menus

Gene name and product information: The description lines are standardized and indexed so that features sharing the same description lines can be retrieved. Access to the nucleotide and amino acid sequences of the feature are also provided.

Basic location information and context map: Clicking on the 'Graphical display in Artemis' open up an Artemis applet – which will be discussed further. The applet allows the feature to be viewed in the context of the sequence and additional annotation, such as UTRs.



Details of protein domains defined by Pfam, Interpro, PRINTS, SMART, PROSITE, TIGRFAM, with links to annotation of these families.

Gene Ontology associations: Links will take you to the descriptions of the terms as well as other proteins annotated to the same ontology node.

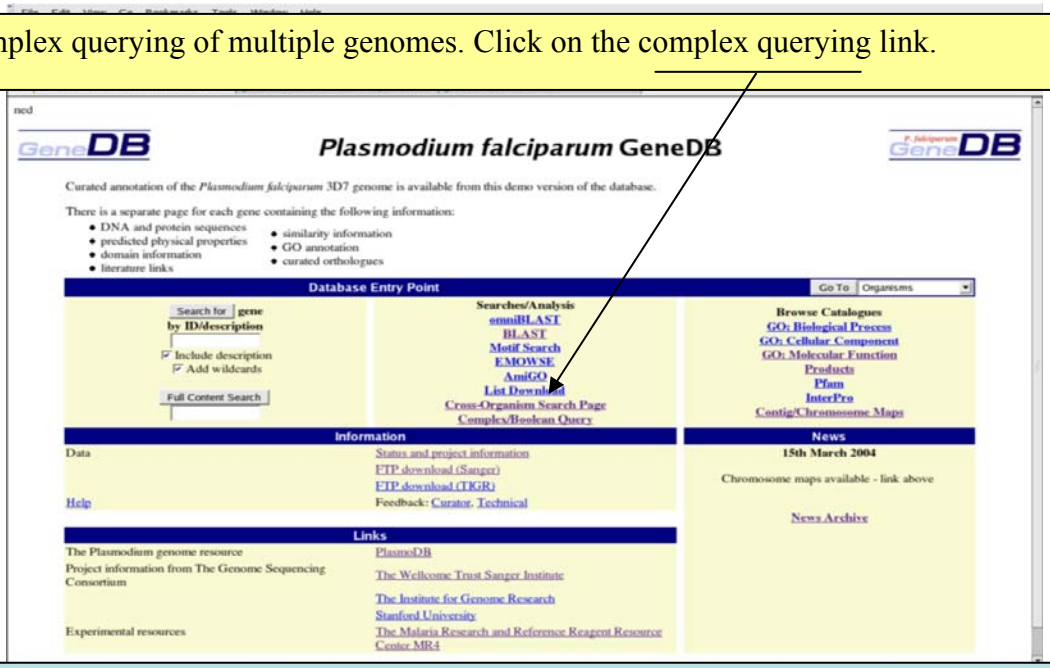
7 ***Reminder***: Return to box 4 to try the other simple keyword searches if you haven't already. Does PFB0210c appear in the results from these searches. Also try entering the search terms in the ***Full Content Search*** box to assess how this affects the results. Use quotes to group two search terms into a phrase, e.g. "hexose transporter" looks for the occcurence of these two terms together.

Through this exercise you should have seen that when searching by geneID/description the gene is found when using, **PfHT**, monosaccharide transporter but not when using sugar transporter, glucose transporter or fructose transporter. **PfHT** is a gene name and because this has been annotated into the database from the literature it is detected in the database. The other search terms are descriptions of the product of the gene and although the protein can transport fructose and hexose it is described as a monosaccharide transporter in the database. This is just an example of how the way in which a gene is described can affect the results of simple searches. It is always better to try several search terms and compare the results. Doing a full content search will search all of the annotation fields associated with a gene. Thus it tends to be comprehensive but give many hits.

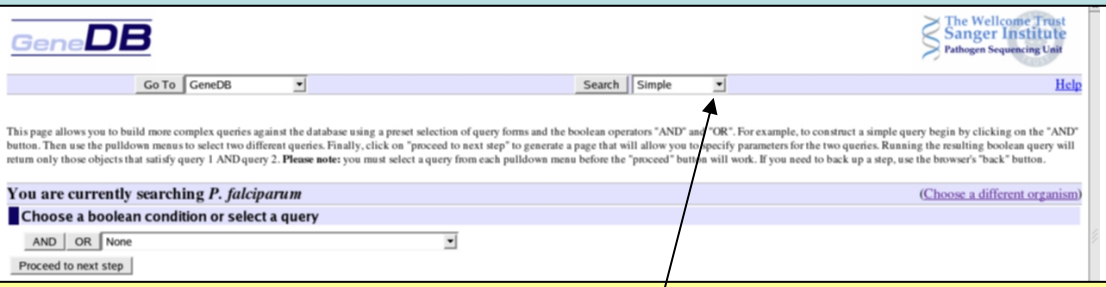
8 We're now going to move onto complex querying which allows searching of several genomes concurrently if desired, and allows a diverse range of queries to be used.

Exercise 2 Searching of multiple genomes using more complex keywords, manipulation of and downloading results sets.

9 Complex querying of multiple genomes. Click on the complex querying link.



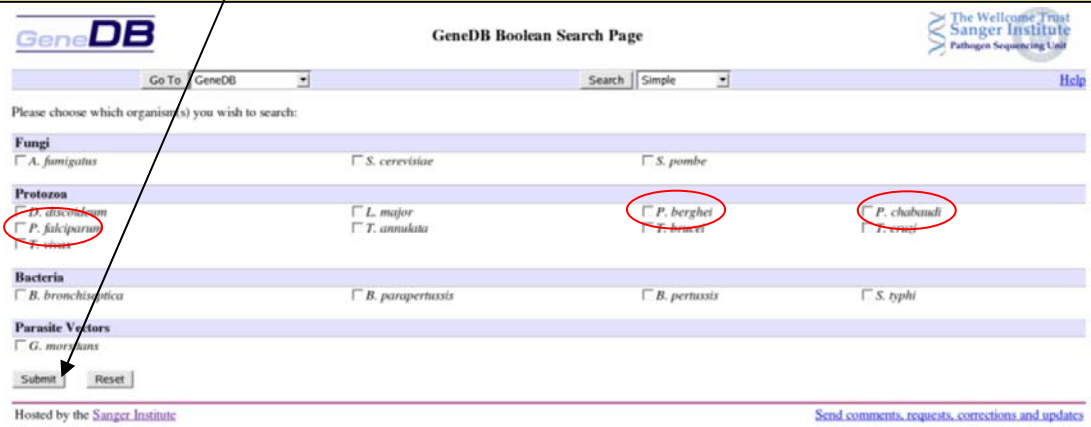
10 This will bring you through to the page below. Since you are going to apply the search to all three Plasmodium genomes in GeneDB, you have to select a **complex** search.



11 Select *complex* from the search drop down menu which will take you through to the following page.

12 Click on the boxes for *P. falciparum*, *P. berghei* and *P. chabaudi* (circled red)

Then click the submit button. This will take you back to the complex search page.



13 Complex queries can be built up using this page. It uses a Boolean approach. Many different data types can be search and **AND** or **OR** can be used to enhance searching.

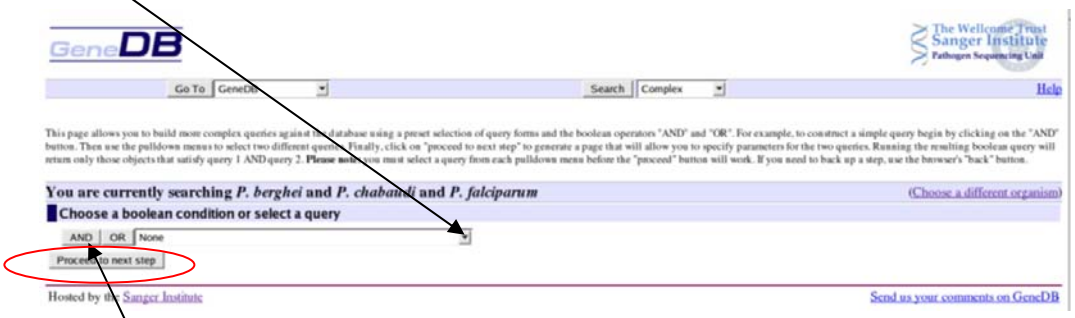
14 Our objective is to find hexose transporters in the three Plasmodium genomes. Try the following complex search strategies and compare the results you obtain. The following pages show you how to set up the first query. Then try numbers 2 and 3 for yourself.

1 (Proteins with a product containing a particular keyword or phrase: *transporter*) **AND** (GO component: *membrane*).

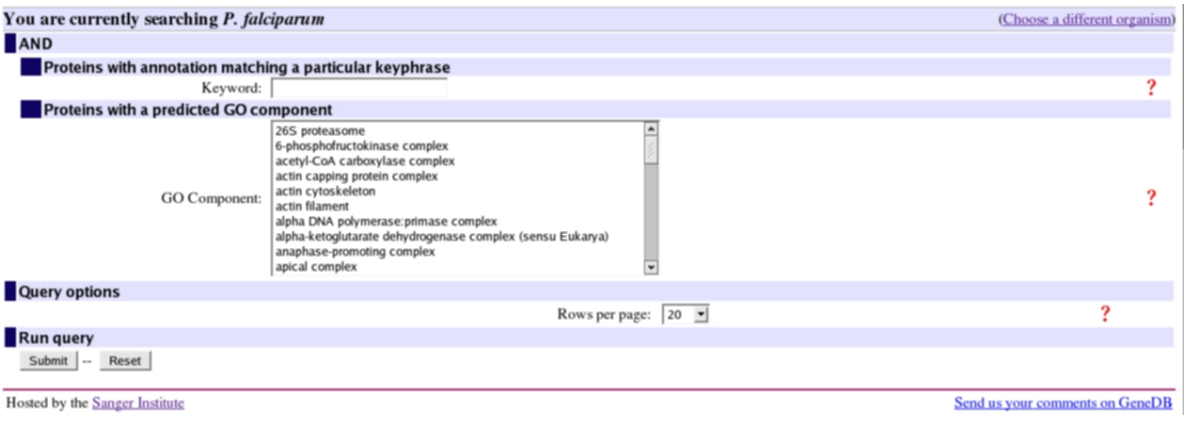
2.(Proteins containing one or more transmembrane domains, try between 8 and 14) **AND** (Proteins with a particular GO process: *transport*).

3.(Proteins with a product containing a particular keyword or phrase: *transporter*) **AND** (Proteins with a signal peptide).

15 Click on the AND button and a second pull down menu should appear below the first. Select *Proteins with a product containing a particular keyword or phrase* by clicking on the first pull-down menu and selecting this option.

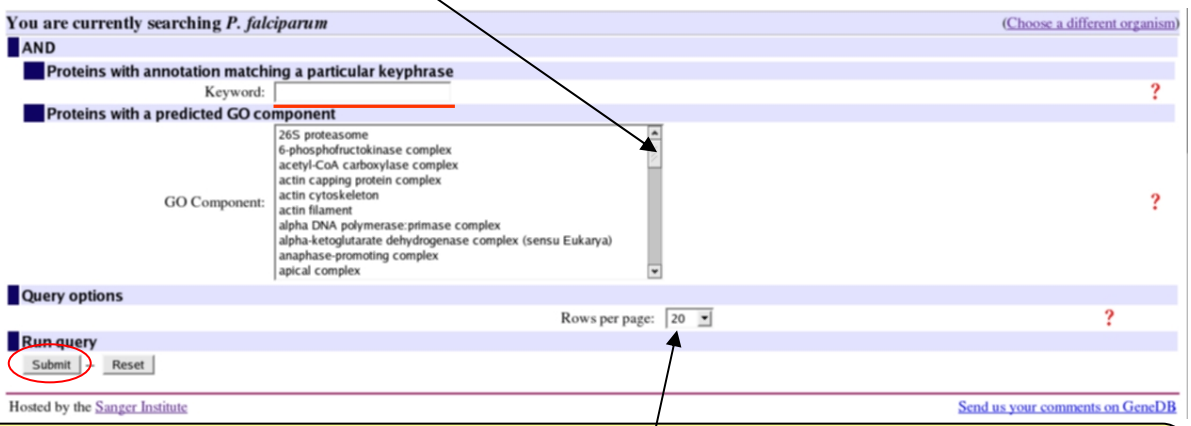


16 In the second pull down menu, select the option *Proteins with a predicted GO component*. Then click on the proceed to next step button (circled red). The screen should appear as below.



17 Type *transport* in the Keyword box (underlined red)

Then select use the scroll bar to scroll down the GO term until you get to membrane and then select it. Then click the submit query button (circled red).



18 Select rows per page to 100 from the drop down menu.

Then select use the scroll bar to scroll down the GO term until you get to membrane and then select it. Then click the submit query button (circled red).

19 The query should return 74 results. Examine the results to see which proteins could be hexose/glucose/monosaccharide transporters and whether PFB0210c is present. Then use the browser back button to go back and submit the second and third queries. Note that the results page tells you whether the protein has been manually or automatically annotated.

20 We're now going to look at ways of treating the results sets that are obtained from our Boolean searches. We can add, subtract and intersect different results sets using the history page. We'll also look at downloading the results sets.

GeneDB — Results 1 - 13 of 13 — The Wellcome Trust Sanger Institute, Pathogen Sequencing Unit

Go To GeneDB — Search Simple

Query: (Genes for *pberghei:pchabaudi:malaria*) intersect ((Proteins with a product containing the keyword or phrase 'transporter') intersect (Proteins containing a predicted signal peptide.))

[1-13]

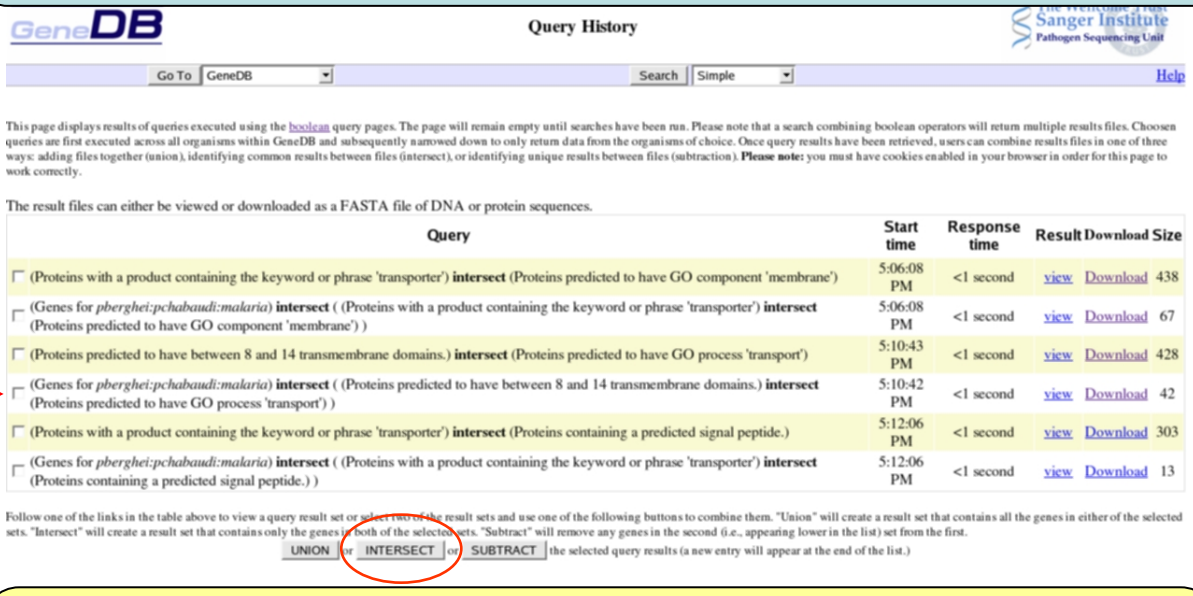
	name	id	organism	description	
1		PB000033.01.0	*P. berghei*	transporter protein, putative	Automatic Annotation
2		PC000560.00.0	*P. chabaudi*	phosphate transporter, putative	Automatic Annotation
3		PC000437.02.0	*P. chabaudi*	transporter protein, putative	Automatic Annotation
4		PB000168.03.0	*P. berghei*	phosphate transporter, putative	Automatic Annotation
5		PC000604.03.0	*P. chabaudi*	sugar transporter, putative	Automatic Annotation
6		PF11_0141	*P. falciparum*	UDP-galactose transporter, putative	Manual Annotation
7		PFF0450c	*P. falciparum*	putative transporter protein	Manual Annotation
8		PF07_0065	*P. falciparum*	zinc transporter, putative	Manual Annotation
9		PF07_0070	*P. falciparum*	transporter/permease protein, putative	Manual Annotation
10		MAL13P1.206	*P. falciparum*	phosphate transporter, putative	Manual Annotation
11		PF14_0133	*P. falciparum*	ATP-dependent transporter, putative	Manual Annotation
12		PFC0125w	*P. falciparum*	ABC transporter, putative	Manual Annotation
13		PFC0875w	*P. falciparum*	transporter, putative	Manual Annotation

To view this result set later, or to download this data visit the history page.

Start a completely new complex query

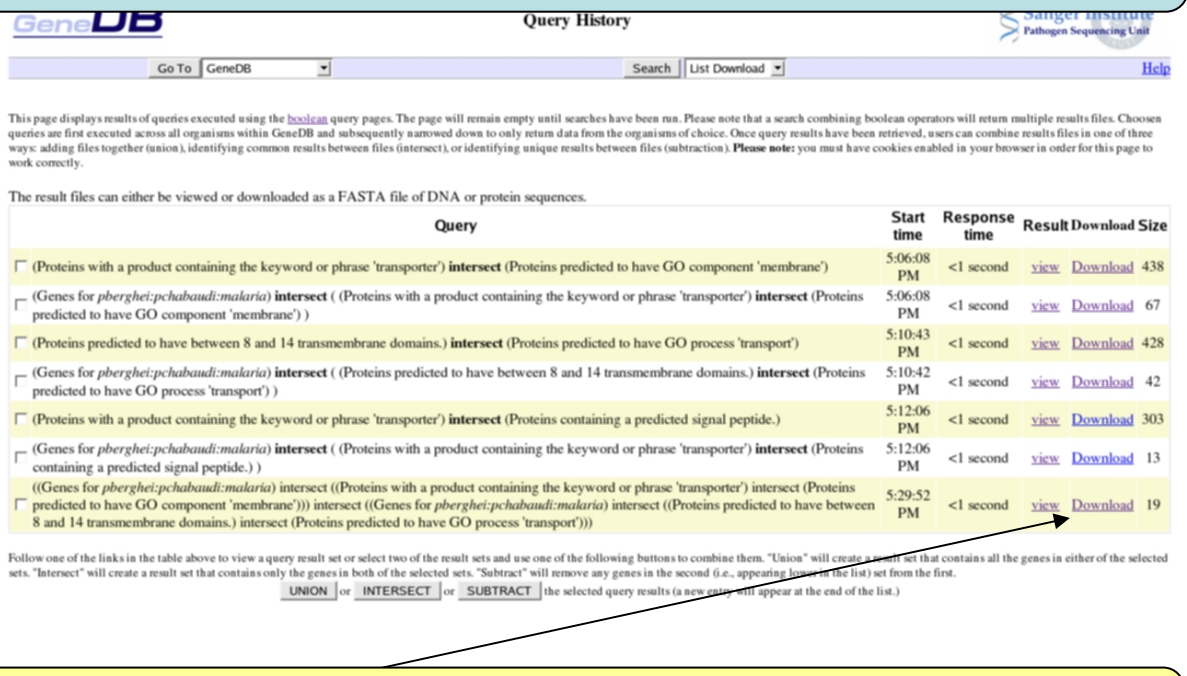
21 Click on the History page link to bring up details of search results.

22 The History page allows results sets to be viewed and downloaded. It is only active for Boolean searches, but is a very useful way of tracking and manipulating results sets. It also allows results sets to be added together (union), the contents of one set removed from another (subtract) and identification of those that appear in both sets (intersect)



23 Click on the boxes for the results sets for query 1 and query 2 (marked by the red arrows) and then click on intersect (circled red). A new results set will appear containing only the genes which occur in the results sets for query 1 and 2. **Why doesn't the gene for the putative monosaccharide transporter in *Plasmodium chabaudi* (PC000736.00.0) appear in the results set for query 2?**

24 We're now going to look at how to download a results set and see what formats and different parts of the dataset can be obtained by choosing different options.



25 Click on the download link for the new results set that you generated by the intersection of queries 1 and 2.

26

In the initial download window the description of the genes in the results list will appear. The lower part of the window boxed is used to select what type of information you want to download. These options include the DNA sequence with introns (unspliced) or without introns (spliced), the protein sequence and either 5' or 3' regions flanking the gene to a chosen number of bases. This would be very useful when examining regulatory elements such as promoter regions or UTR (untranslated region)

27

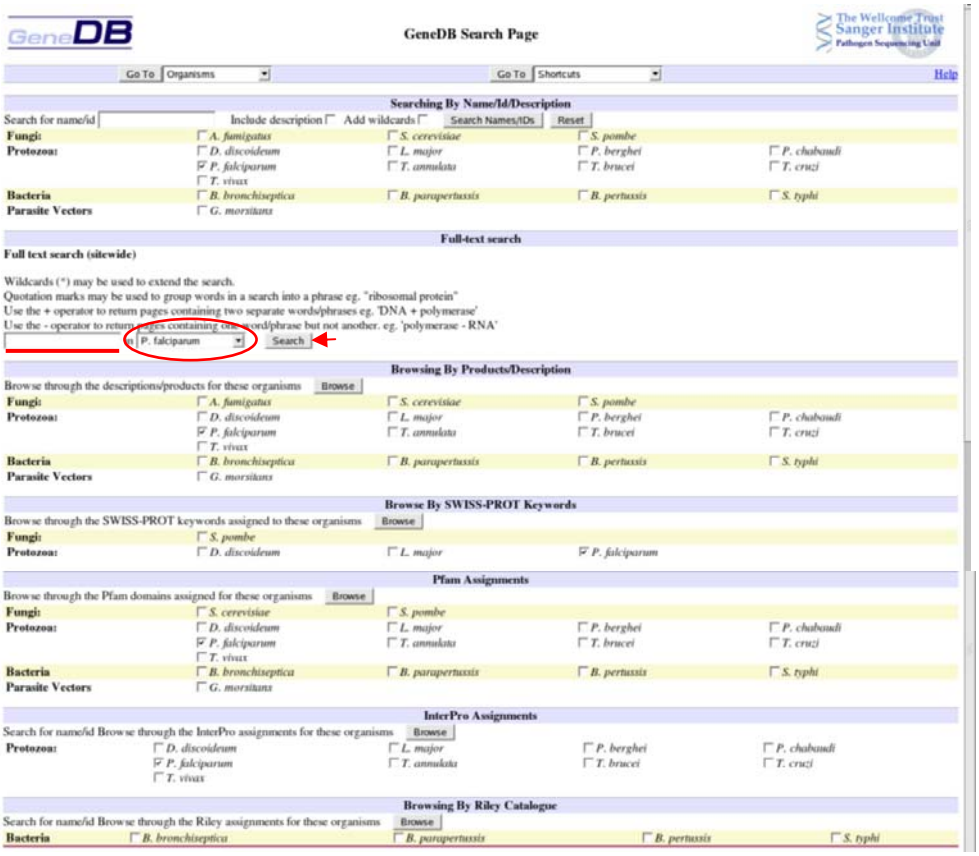
For the sake of our exercise lets assume that we will proceed to compare the amino acid sequences of our transporters. Thus we would click on Protein sequence (arrow) and then on the submit query button (circled). Then to download this information you would save the page from your browser. This is not described here as the process will be particular to the internet browser that your using.

Exercise 3 Search strategies using omniBLAST and browsing of the Pfam domain catalogue

28 We are now going to return to consider and run a few other search strategies which make use of the strengths of GeneDB.

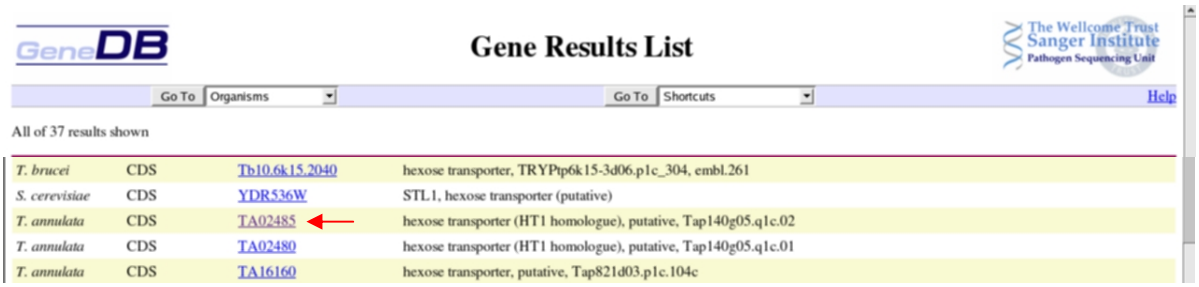
1. Use of a text keyword search across several organisms using the cross-organism search page. This can be a quick and powerful way to identify genes/proteins in other organisms that perform very similar functions to your gene of interest. This can be achieved with simple keywords and requires little previous knowledge about the gene of interest. Once a gene or protein has been found that meets the keyword criteria, e.g sugar+transporter, the DNA or Protein sequence can be searched against any genome in GeneDB using omniBLAST. In our case we would want to search the three Plasmodium genomes.
2. Another powerful approach makes use of the fact that many protein domains that are diagnostic of a particular function have already been characterised and assigned to many genes within the database. Thus, if we know that our gene of interest has a particular Pfam or Interpro domain then we can browse through the Pfam or Interpro catalogue for genes which have this domain. This can be done concurrently for several organisms using the Cross-Organism search page.

29 Select the Cross-Organism Search Page. The link can be found on the *Plasmodium falciparum* main page or any of the organism main pages. See the figure under **box 4.** In the full text search section (underlined red) enter *"hexose transporter"+"glucose transporter"* (include the quotes). In the adjacent box circled red select ***All organisms*** from the pull down menu. Then click on the adjacent Search button. (red arrow)

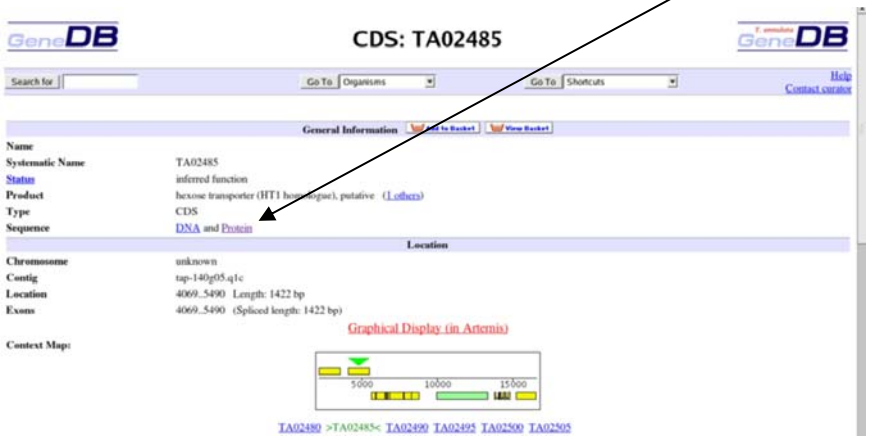


30 A results list will then be displayed as below (note that the whole list is not shown). We are now going to take the protein sequences of one of these genes and search for similar protein sequences in the genomes of P. *falciparum*, *P. berghei* and P. *chabaudi* using Omniblast. This uses the wublastp algorithm. The example below will take you through this process for one protein, but if you have time try one or two others. To start with use the sequence of a protein from *T. annulata* (TA02485) as this as been annotated as a hexose transporter homologue (HT1 homologue) which is the name of the gene in *P. falciparum* as it appears in the literature, and so it is likely that it is an orthologue of the protein in *P. falciparum*.

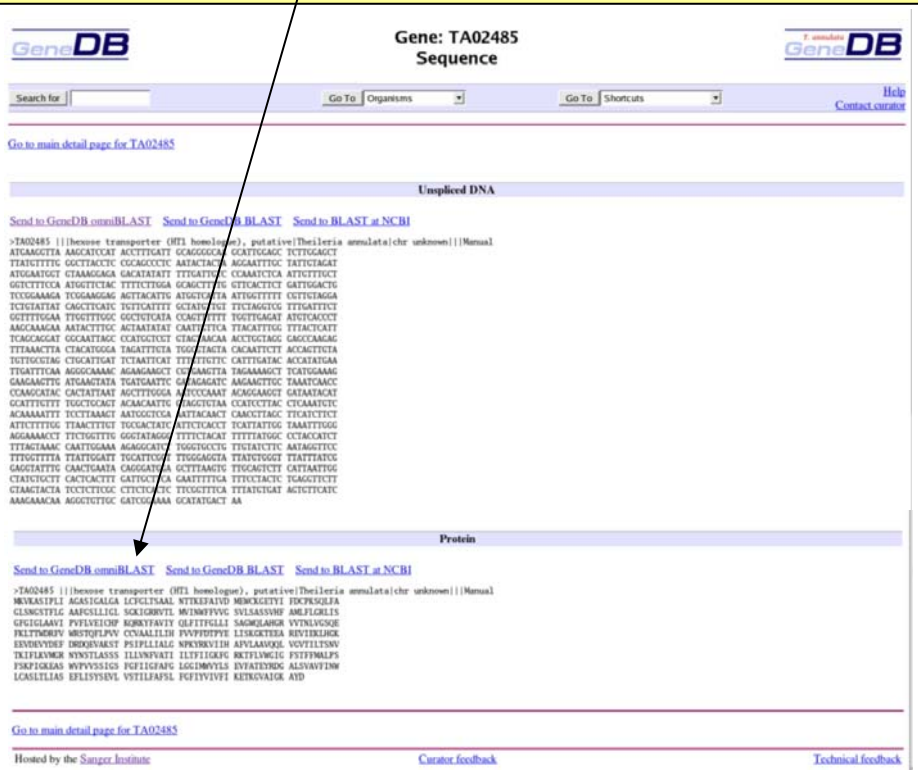
31 Click on the link for as shown below. This will take you through to the feature page for this gene.



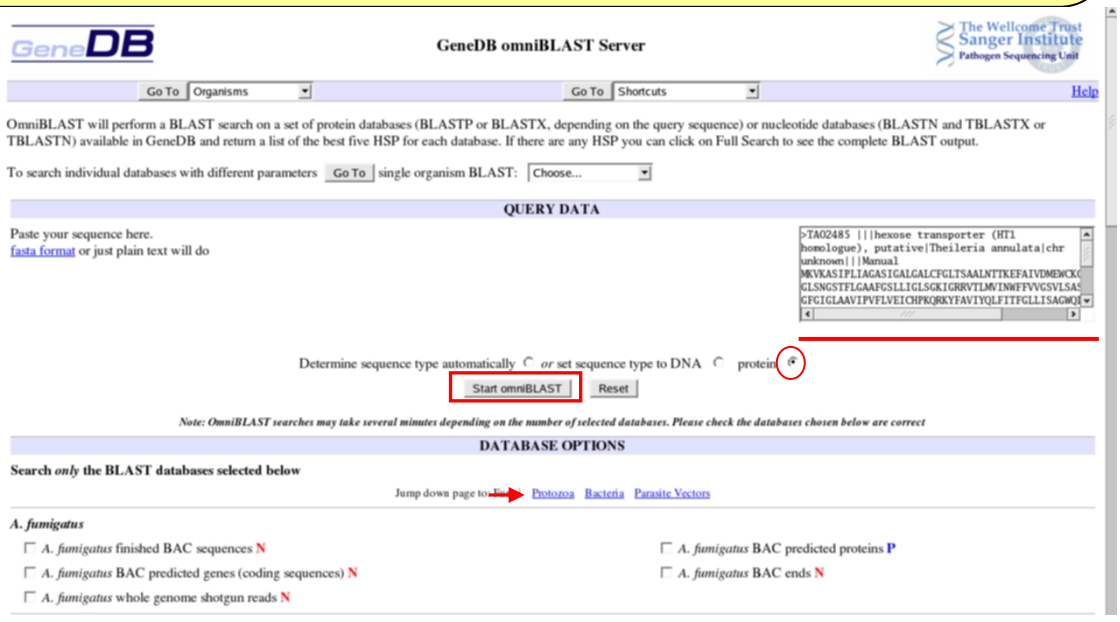
32 To access the protein sequence to do an OmniBlast search click on the protein link.



33 Click on the Send to GeneDB omniBLAST link to send this sequence as a query to the GeneDB omniBLAST server.

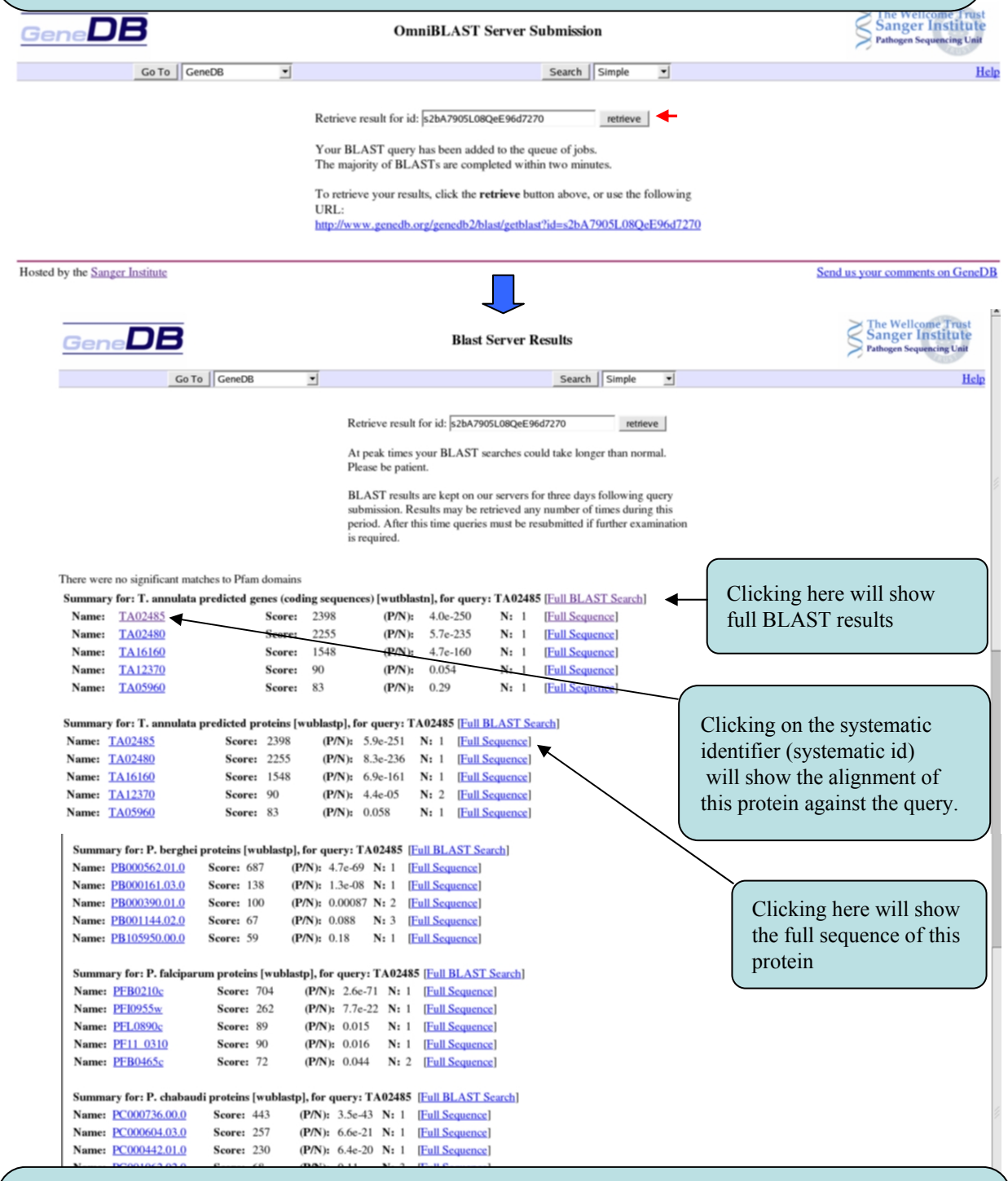


34 The sequence will automatically be put into the query box (red underlined) and the protein option selected (red circled). Choose the genomes that you want the query to be searched against (note that only the top part of the page is shown). To go down to protozoan genomes click on the jump down to Protozoa link (red arrow). Note that the *T. annulata* sequence will automatically be selected. Select the protein databases for se *P.falciparum*, *P. chabaudi* and *P.berghei* by clicking on the small text box adjacent to **P**, then click on the start omniBLAST button (red box).



35 Note that OmniBLAST can be used to search on the basis of DNA sequence also. Sequence can also be pasted into the query box (FASTA or plain text) and searched.

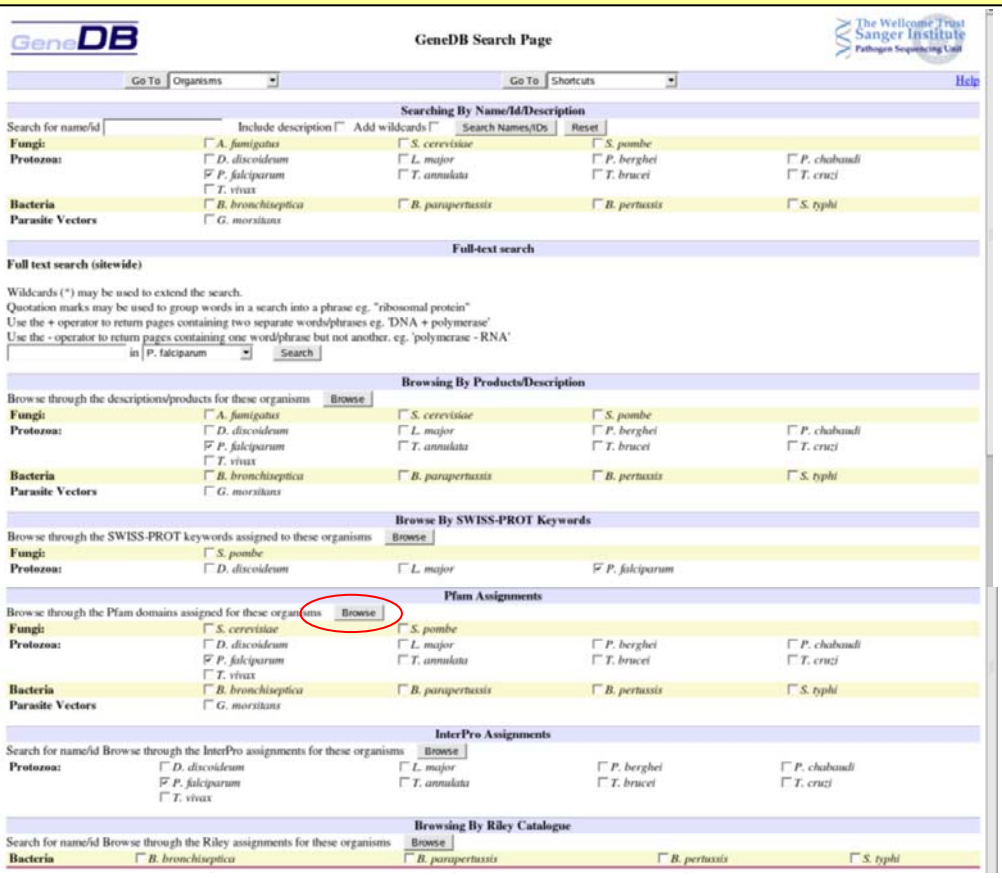
36 Click on the Retrieve button (red arrow). The omniBLAST search may take a while depending on the number size of the search. Once completed the omniBLAST results are presented in a summarised format (as shown below) as BLAST output files are large and detailed. The top five hits in each search are summarised. **Does this search detect PFB0210c (PfHT)? Are these results consistent with what you thought were the orthologues of pfHT in *Plasmodium berghei* and *Plasmodium chabaudi* from your previous searches?** Once you have looked at the search results try clicking on the various options in the results page.



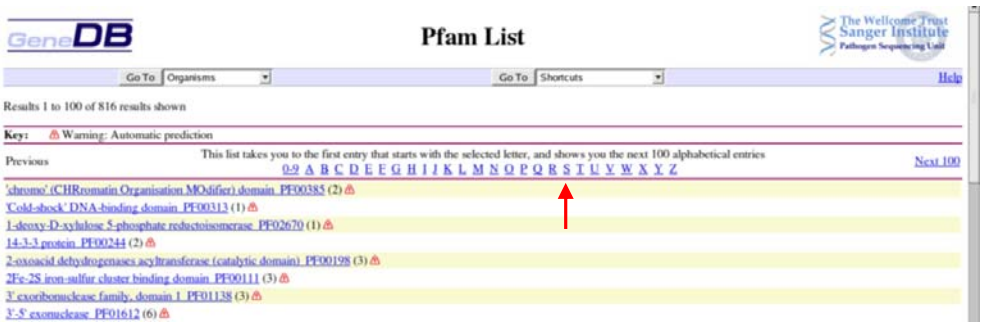
37 This approach has demonstrated how an omniBLAST search can identify the gene of interest in your organism when a well-annotated orthologue exists in another organism. So this is a useful alternative strategy to searching on keywords alone, which we have seen can in some cases be misleading. It also shows that the full text search (site wide) is a powerful way of searching the annotation of all the genomes in GeneDB for possible orthologues.

38 An alternative way of identifying potential orthologues is the presence of a protein domain that is associated with that function. This approach also makes use of the Cross Organism search Page which allows browsing of Pfam and Interpro assignments across several genomes concurrently. Let us assume that our previous searches had uncovered that the Pfam domain *Sugar (and other) transporter (PF00083)* is a Pfam domain associated with Hexose Transporters. Note that in PF00083, PF stands for Pfam. To view more details about this protein domain goto the gene page (see under box 6 for reference) and click on PF00083 which is in a red box in the figure under box 6. We want to search for proteins in *Plasmodium falciparum*, *Plasmodium chabaudi*, *Plasmodium berghei* that also have this domain.

39 Go to the Cross-Organism search page (for reference see figures under Box 9 and Box 2 if needed). In the Pfam Assignments section click the boxes for *P. falciparum*, *P. chabaudi*, *P berghei* and then Browse button (circled red).



40 In the Pfam list click on 'S' (red arrow) and then select sugar (and other) transporter from the list.



41

You should obtain a results set something like that below. Does the set contain the orthologues for PfHT in *P. berghei* and *P. chabaudi*?

GeneDB

Search results

The Wellcome Trust Sanger Institute
Pathogen Sequencing Unit

Go To Organisms | Go To Shortcuts | Help

All of 11 results shown

Report Download

P. falciparum	CDS	PFI0785c	transporter, putative
P. falciparum	CDS	PFB0275w	hypothetical protein, conserved
P. chabaudi	CDS	PC000604.03.0	sugar transporter, putative
P. chabaudi	CDS	PC000855.00.0	transporter, putative
P. chabaudi	CDS	PC000736.00.0	monosaccharide transporter, putative
P. berghei	CDS	PB001031.03.0	conserved hypothetical protein
P. falciparum	CDS	PFB0210c	monosaccharide transporter, putative
P. chabaudi	CDS	PC000438.02.0	conserved hypothetical protein
P. berghei	CDS	PB000562.01.0	monosaccharide transporter, putative
P. falciparum	CDS	PFI0955w	sugar transporter, putative
P. berghei	CDS	PB000243.00.0	transporter, putative

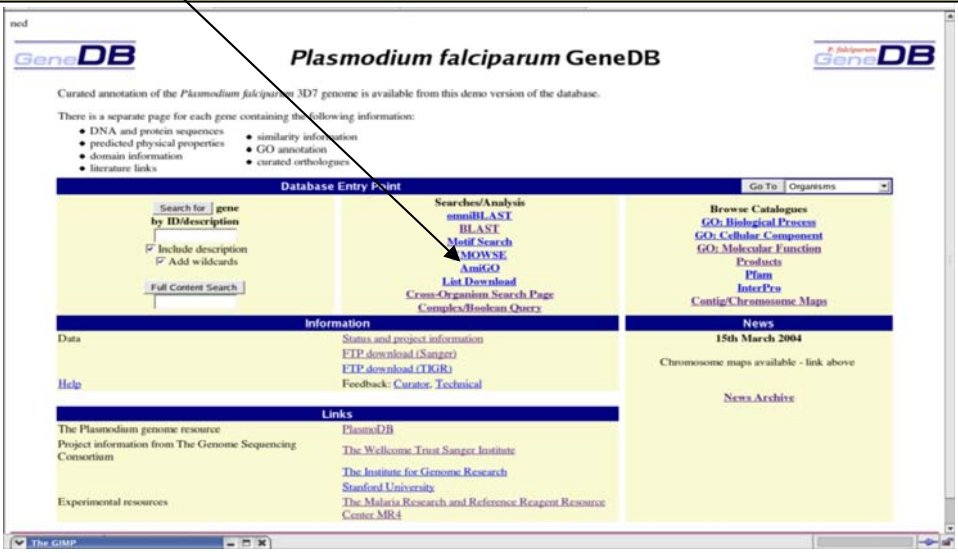
Hosted by the Sanger Institute

Send comments, requests, corrections and updates

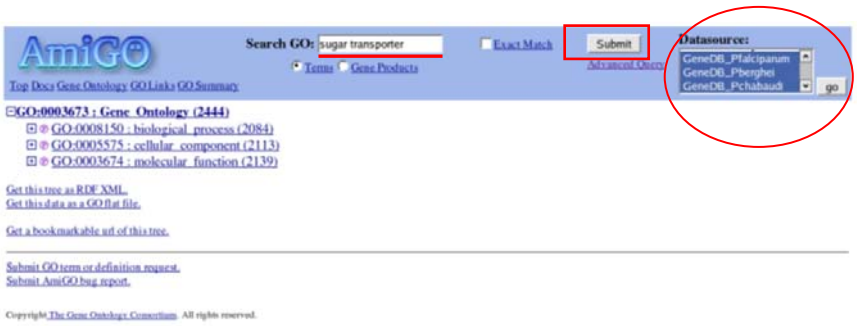
Exercise 4 Searching using Gene Ontology annotation

42 Another search strategy is to search on the basis of Gene Ontology terms. Gene Ontologies are structured vocabularies that are designed to describe biological processes in an accurate and consistent way (for more information see http://www.geneontology.org.) The ontology is composed of three terms the molecular function, biological process and cellular component (location) of a protein. Where evidence exists from the literature, from sequence analysis or other sources Gene Ontology terms for function, process and component are attributed to that gene. AmiGO is a database of Gene Ontology associations that is designed and maintained by the Gene Ontology consortium. It allows searching and browsing of gene ontology annotation across many genomes (including those which are not annotated and curated for GeneDB) and is accessible via GeneDB. It can be a powerful way to search for genes with similar function across several organisms: in our case the search for transporters of glucose and hexose. The example below shows how to set up this query. One you've tried it and have become familiar with it, try some of the other suggested searches or perhaps one that would be of interest to your own research.

43 Click on the AmiGO link to go to the AmiGO simple query search page.



44 Enter sugar transporter in the Search GO box (underlined red). In the datasource box (circled red) scroll down and click on GeneDB_Pfalciparum, GeneDB_Pbergei and GeneDB_Pchabaudi (you'll need to press the shift button whilst clicking to select all three). Then click on the submit button (boxed red).



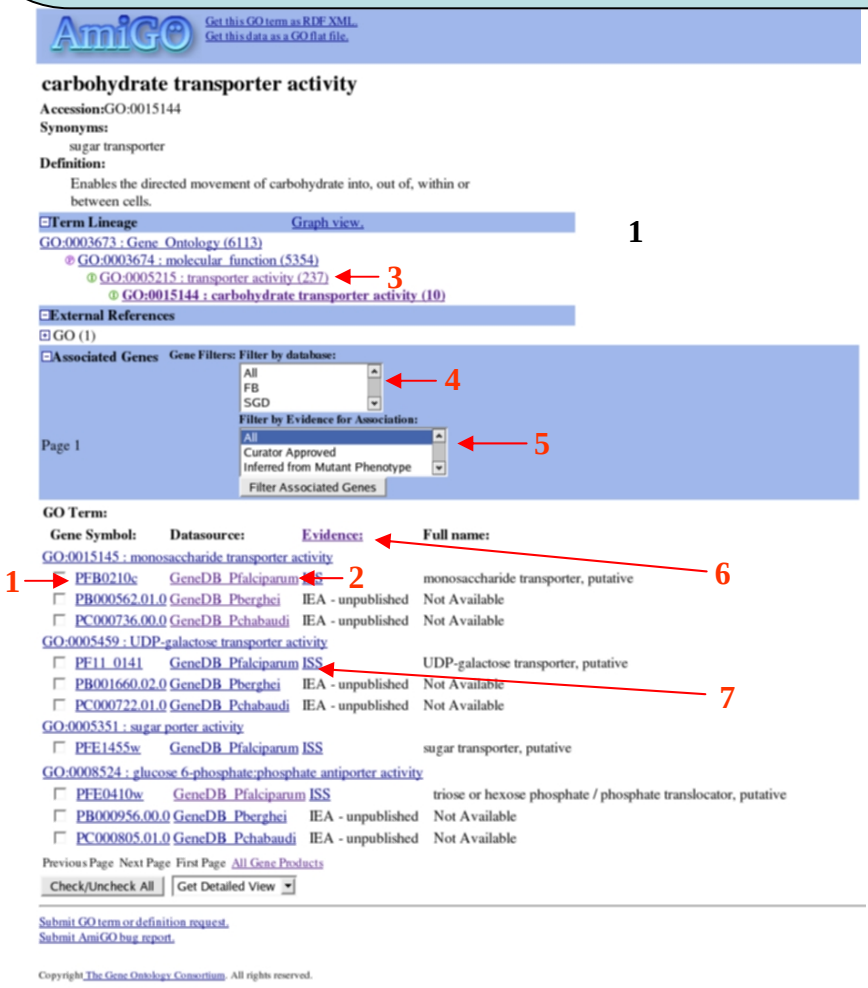
45

The results should be similar to those below – how do the results compare to your previous searches? Is this search successful in identifying the putative orthologues of PfHT in *P. chabaudi* and *P. berghei*? Clicking on the gene name (red arrow 1) will display the exact GO ontology annotation for that gene.

To go the GeneDB gene page click on the name of the database e.g. GeneDB_Pfalciparum (red arrow 2).

The Term Lineage section shows that "carbohydrate transporter" term is a subset or "child" of transporter actvity. Clicking on *GO0005125: transporter activity* (arrow 3)will display all genes with transporter activity within *P falciparum*, *P. berghei* and *P. chabaudi*.

The Associated Genes section can be used to apply the same search to other organism databases (arrow 4) and also to filter the results by evidence code (arrow 5). The evidence code provides information on the type of data that was used to apply a particular GO term to that gene. ISS is Inferred from Sequence or Structural similarity and is used when similarites such as BLAST hits, the presence of protein domains or other features based on sequence or structural similarity. IEA is inferred from Electronic Annotation and is used when similarities have been transferred from automated annotation and have not been reviewed by a curator. For a more detailed description of evidence codes click on the evidence link (arrow 6). If the evidence code has a link this will provide more information about the evidence for the GO term (arrow 7).



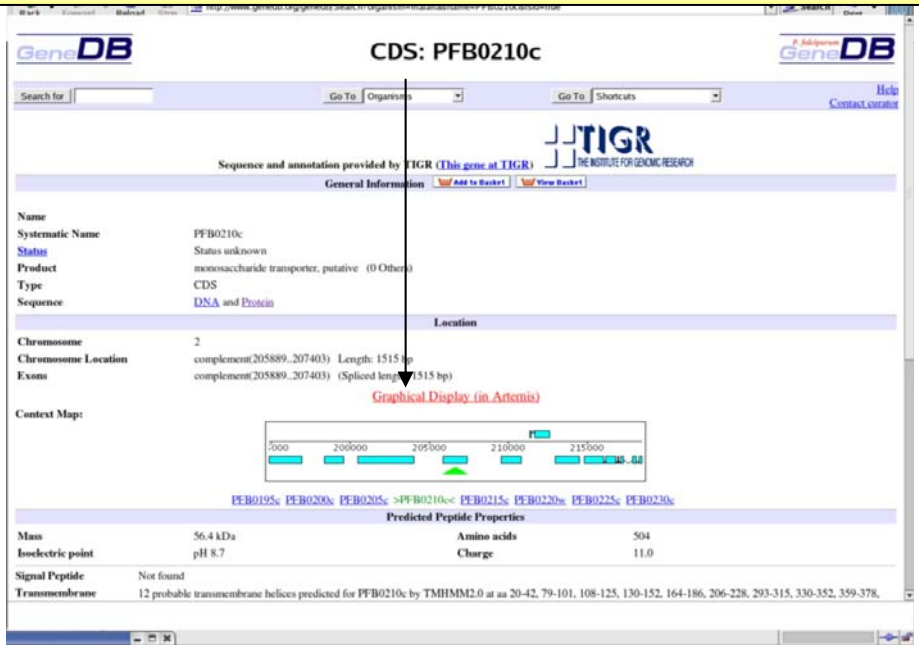
46

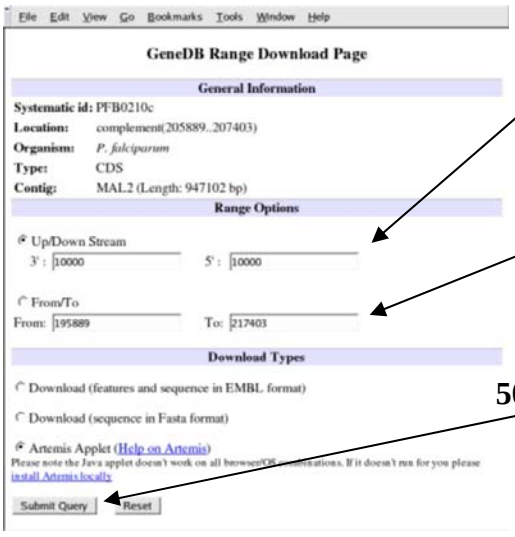
Hopefully these exercises have familiarised you with several strategies for data mining in GeneDB, and given you ideas how GeneDB could be applied to your own research area. If you have any further questions please ask a demonstrator, or after the course please address your queries to the GeneDB team who will be happy to help you. See box 60 and the figure below for details of email links.

Exercise 5 Use of the Artemis Applet

47 We are now going to look at the use of the artemis applet. It will only be considered briefly since you have already covered the use of Artemis. It can be launched from within GeneDB and is a useful way of viewing the gene in the context of the genome. It is especially useful for visualising intergenic regions, promoters, 5' and 3' untranslated regions, intron-exon boundaries, as well as many other features.

48 GO to the gene page for PFB0210c. You can do this is many ways, but one is to go to the *Plasmodium falciparum* page in GeneDB (see box 3). Then enter **PFB0210c** in the *Search for gene by ID/ description* box (see box 4). Then from the gene page click on the **Graphical display in artemis link**





49 This window allows you to specify the region that is opened by the viewer. The default is 10 Kb upstream and 10 Kb downstream of the gene selected. This can be modified by the user.

You can also selected the region that you want to view using coordinates.

50 For the purposes of this example we will use the default settings. Click on the submit query button

52

The Artemis Applet retains nearly all of the functions that is has if it is run locally. Refer to your notes from module on Artemis if necessary. There are far to many functions to describe them all here so were going to look at a few which are relevant to our investigation.

The hexose transporter that we are looking (PfHT, systematic identifier PFB0210c) has been characterised biochemically. It is able to transport glucose and fructose down a chemiosmotic gradient as a classic uniporter. Some residues that define substrate specifity have been identified by mutagenesis experiments (Woodrow et al., 2000). If the Glutamine residue at position 169 is changed to Asparaginine, a mutation denoted by (Q169N), the ability to transport fructose is abolished, but the ability to transport glucose is retained. This residue exist within a the 5 th predicted transmembrane helix.
We'll use the artimis applet to:
• Look at the annotation for pfHT-1 (systematic identifier PFB0210c)
•View hydrophobicity/hydrophilicity plots for the protein
• Examine the amino acid sequence around position 169

Use these scroll bars to adjust the DNA and protein views

53

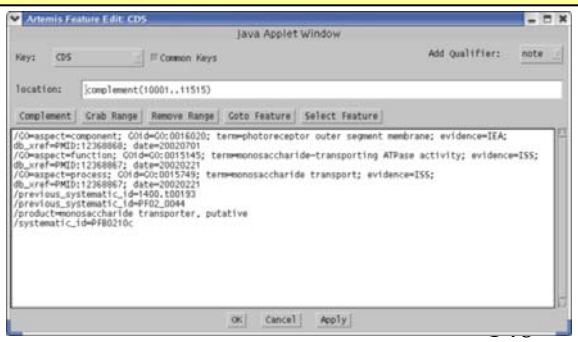
Click here to select a gene and press E (or ctrl E) to view the annotation for the gene.

Find PFB0210C and view the annotation for this gene. It should be something like that shown below.

Note the genes may be coloured due to updates in annotation

54

Examine the annotation briefly before closing this window.
Select the gene of interest by **double** clicking on its entry in the gene list (green arrow) or on its box in the DNA or Protein view. A solid black line will appear around the gene in the DNA and protein view.

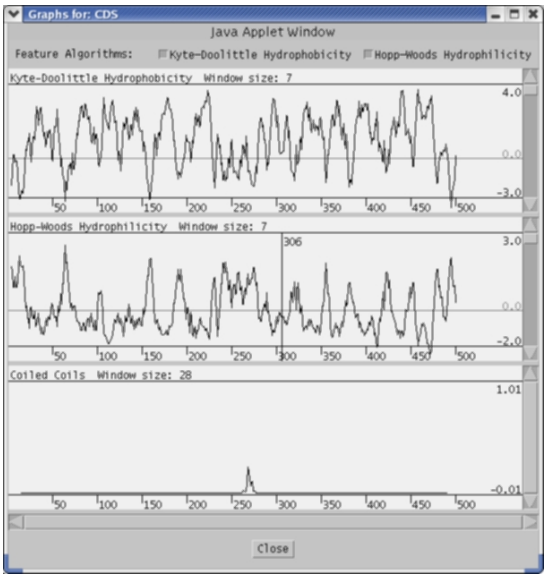


55 While the gene is select click on the *View* menu. Then from this drop down menu choose *show feature plots* (the bottom one in the list)

Note: the genes in the display may appear coloured due to updates in annotation

56 These plots show hydrophobicity (upper) and hydrophilicity (lower).

Is residue 169 located within a hydrophilic or hydrophonic region of the protein? (note you can click within this diagram to get a line from the x-axis up to the curve)



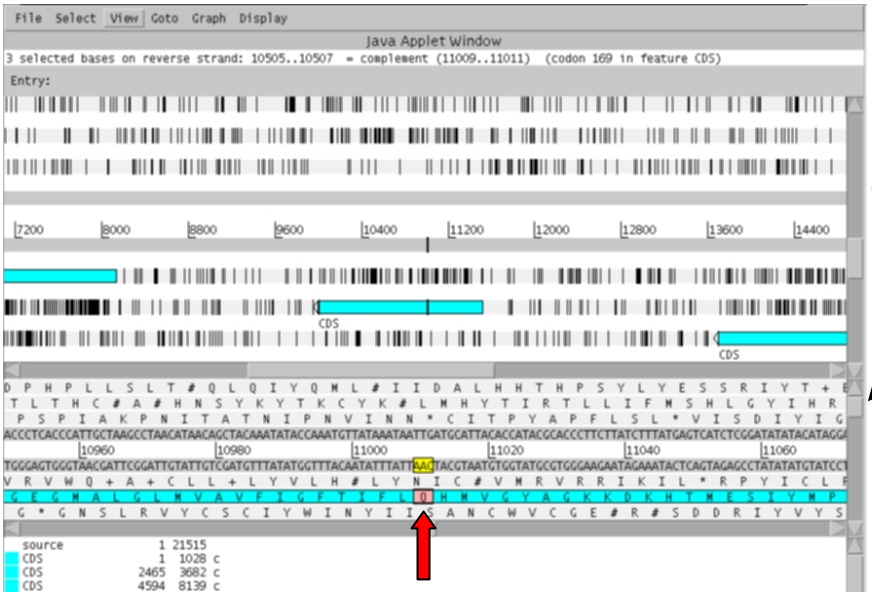
57 Close this window. You should return to the Artemis window. Select the gene of interest again if it is not already selected.

58 Now we are going to use the Artemis applet to find a specific amino acid, Glutamine Q at position 169, within the selected protein.

59 Click on *Goto* in the menu bar and select *Feature amino acid..*

A box will appear, enter 169 and click on OK.

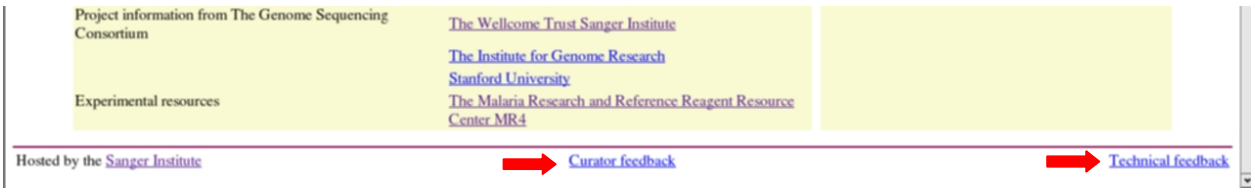
The Glutamine residue will be highlighted in the protein view.



60 If necessary use this scroll bar to adjust the protein view so that you can see individual amino acids

61 This part of the exercise has made you aware that you run use the Artemis applet from within GeneDB to view the detailed structure of the gene within its genomic content.

Annotation is at different stages for different genomes, and is actively improved where genes are manually curated. There will be occasions where the annotation may be misleading, incomplete or not as comprehensive as it could be. If you have any comments or about a particular gene's annotation, or can provide data that you think improves the knowledge base, then don't hesitate to contact the curator for that organism via email. Responses are normally provided within one or two working days. If your query or suggestion is of a technical nature, or is something that could apply to the whole of GeneDB, rather than the annotation of a particular gene or organism, then it may be better addressed to technical feedback. There are links on the bottom of each page in GeneDB (see image below).



References

Rice,P. Longden,I. and Bleasby,A. **(2000)** *Trends in Genetics* **16** (6) *276-277*
EMBOSS: The European Molecular Biology Open Software Suite

Carver T.J., Mullan L.J., **(2002)** *Comparative and Functional Genomics* **3** (1) *75-78*,
A new graphical user interface to EMBOSS

Rutherford *et al.***(2000)** *Bioinformatics* **16** (10) *944-945*
Artemis: sequence visualization and annotation

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrel, B.G. and Parkhill, J. **(2005)** *Bioinformatics* **21** (16) *3422-3423* ACT: the Artemis comparison tool.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape. (1997) Pathogenicity islands of virulent bacteria:structure, function and impact on microbial evolution. *Mol Microbiol* **23:;** 1089-97.

Appendices

Appendix I: Artemis minimum hardware and software requirements.

Artemis and ACT will, in general, work well on any standard modern machine and with most common operating systems. It is currently used on many different varieties of UNIX and Linux systems as well as Apple Macintosh and Microsoft Windows systems.
Note that the ability to run external programs (such as BLAST and FASTA) from within Artemis and ACT is available only on UNIX and Linux systems.
Minimum memory requirements for people working on whole genomes are approximately 128 megabytes for Artemis and 128 megabytes per genome for ACT. Analysis of cosmid sized sequences can comfortably be achieved with less memory.

Appendix II: ACT comparison files

ACT supports three different comparison file formats:

1) BLAST version 2.2.2 output: The blastall command must be run with the -m 8 flag which generates one line of information per HSP.
2) MEGABLAST output: ACT can also read the output of MEGABLAST, which is part of the NCBI blast distribution.
3) MSPcrunch output: MSPcrunch is program for UNIX and GNU/Linux systems which can post-process BLAST version 1 output into an easier to read format. ACT can only read MSPcrunch output with the -d flag.

Here is an example of an ACT readable comparison file generated by MSPcrunch -d.

```
1399 97.00 940 2539 sequence1.dna 1 1596 AF140550.seq
1033 93.00 9041 10501 sequence1.dna 9420 10880 AF140550.seq
828 95.00 6823 7890 sequence1.dna 7211 8276 AF140550.seq
773 94.00 2837 3841 sequence1.dna 2338 3342 AF140550.seq
```

The columns have the following meanings (in order): score, percent identity, match start in the query sequence, match end in the query sequence, query sequence name, subject sequence start, subject sequence end, subject sequence name.

The columns should be separated by single spaces.

Appendix III: Feature Keys and Qualifiers – a brief explanation of what they are and a sample of the one's we use.

1 – Feature Keys: They describe features with DNA coordinates and once marked, they all appear in the Artemis main window. The ones we use are:

➔ CDS: Marks the extent of the coding sequence.
➔ RBS: Ribosomal binding site
➔ misc_feature: Miscellaneous feature in the DNA
➔ rRNA: Ribosomal RNA
➔ repeat_region
➔ repeat_unit
➔ stem_loop
➔ tRNA: Transfer RNA

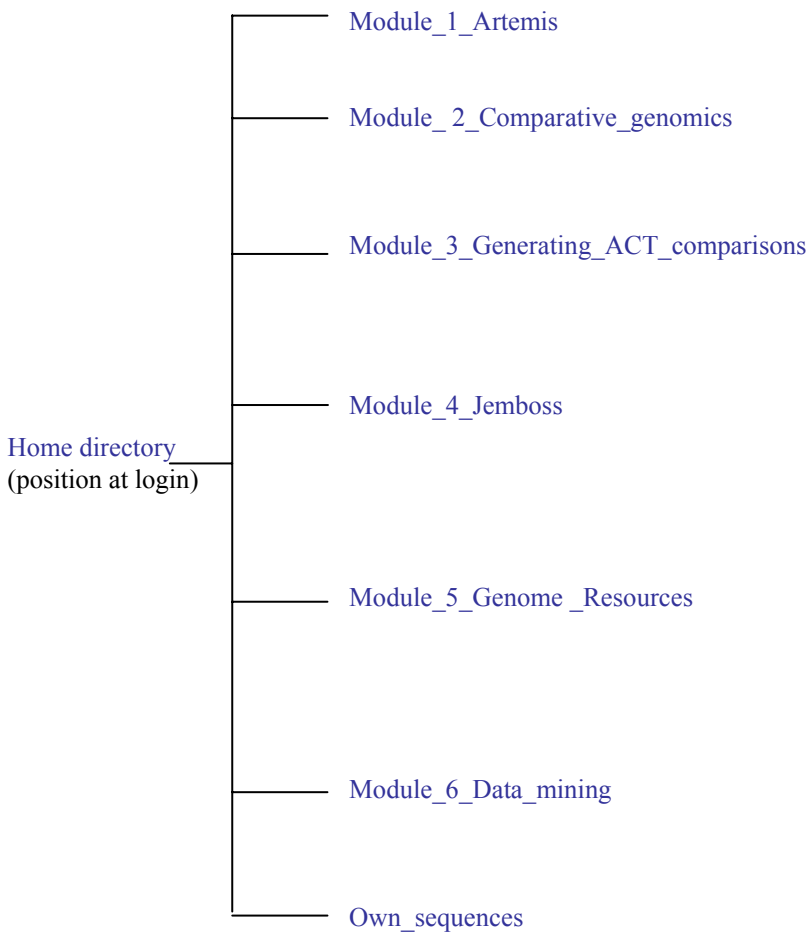
2 – Qualifiers: They describe features with protein coordinates. Once marked they appear in the lower part of the Artemis window. They describe the gene whose coordinates appear in the 'location' part of the editing window. The ones we commonly use for annotation at the Sanger Institute are:

➔ Class: Classification scheme we use "in-house" developed from Monica Riley's MultiFun assignments (see Appendix VI).
➔ Colour: Also used in-house in order to differentiate between different types of genes and other features.
➔ Gene: This qualifier either gives the gene a name or a systematic gene number.
➔ Label: Allows you to label a gene/feature in the main view panel.
➔ Note: This qualifier allows for the inclusion of free text. This could be a description of the evidence supporting the functional prediction or other notable features/information which cannot be described using other qualifiers.
➔Partial: When a region in the DNA hits a protein in the database but lacks start and/or stop codons and the match does not include the whole length of the protein, it can be considered as a partial gene.
➔ Product: The assigned possible function for the protein goes here.
➔ Pseudo: Matches in different frames to consecutive segments of the same protein in the databases can be linked or joined as one and edited in one window. They are marked as pseudogenes. They are normally not functional and are considered to have been mutated.

The list of keys and qualifiers accepted by EMBL in sequence/annotation submission files are list at the following web page:
http://www3.ebi.ac.uk/Services/WebFeat/

Appendix IV: Schematic of workshop files and directories

Key:
Directories and subdirectories



Appendix V: Useful Web addresses

Major Public Sequence Repositories

DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl.html
Genomes at the EBI	http://www.ebi.ac.uk/genomes/
GenBank	http://www.ncbi.nlm.nih.gov/

Microbial Genome Databases Resources

Sanger Microbial Genomes	http://www.sanger.ac.uk/Projects/Microbes/
TIGR Microbial Database	http://www.tigr.org/tdb/mdb/mdbcomplete.html
Institute Pasteur GenoList databases *Including: SubtiList, Colbri, TubercuList, Leproma, PyloriGene, MypuList, ListiList, CandidaDB,*	http://genolist.pasteur.fr
Pseudomonas Genome Database	http://www.pseudomonas.com/
Clusters of Orthologous Groups of proteins (COGs)	http://www.ncbi.nlm.nih.gov/COG/
SCODBII (*S. coelicolor* database)	http://www.jiio16.jic.bbsrc.ac.uk/S.coelicolor

Protein Motif Databases

Prosite	http://www.expasy.ch/prosite/
Pfam	http://www.sanger.ac.uk/Software/Pfam/index.shtml
BLOCKS	http://blocks.fhcrc.org
InterPro	http://www.ebi.ac.uk/interpro/
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
SMART	http://smart.embl-heidelberg.de
InterPro	http://www.ebi.ac.uk/interpro/index.html

Protein feature prediction tools

TMHMM Prediction of transmembrane helices in proteins	http://www.cbs.dtu.dk/services/TMHMM-2.0/
SignalP Prediction Server	http://www.cbs.dtu.dk/services/SignalP/
PSORT protein prediction	http://psort.ims.u-tokyo.ac.jp/form.html

Metabolic Pathways and Cellular Regulation

EcoCyc	http://ecocyc.org/
ENZYME	http://www.expasy.ch/enzyme/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg
MetaCyc	http://ecocyc.org/

Miscellaneous sites

NCBI BLAST website	http://www.ncbi.nlm.nih.gov/BLAST/
The tmRNA website	http://www.indiana.edu/~tmrna/
tRNAscan-SE Search Server	http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Codon usage database	http://www.kazusa.or.jp/codon/
RNAgenie RNA gene prediction	http://rnagene.lbl.gov/
GO Gene Ontology Consortium	http://www.geneontology.org/
Artemis homepage	http://www.sanger.ac.uk/Software/Artemis/
ACT homepage	http://www.sanger.ac.uk/Software/ACT/
Glimmer	http://www.tigr.org/software/glimmer/
Orpheus	http://pedant.gsf.de/orpheus

Appendix VI: Prokaryotic Protein Classification Scheme used within the PSU

This scheme was adapted for in-house use from the Monica Riley's protein classification <http://genprotec.mbl.edu/riley-lab.html>).
More classes can be added depending on the microorganism that is being annotated (e.g secondary metabolites, sigma factors (ECF or non-ECF), etc).

0.0.0 Unknown function, no known homologs
0.0.1 Conserved in Escherichia coli
0.0.2 Conserved in organism other than Escherichia coli

1.0.0 Cell processes
- 1.1.1 Chemotaxis and mobility
- 1.2.1 Chromosome replication
- 1.3.1 Chaperones

1.4.0 Protection responses
- 1.4.1 Cell killing
- 1.4.2 Detoxification
- 1.4.3 Drug/analog sensitivity
- 1.4.4 Radiation sensitivity

1.5.0 Transport/binding proteins
- 1.5.1 Amino acids and amines
- 1.5.2 Cations
- 1.5.3 Carbohydrates, organic acids and alcohols
- 1.5.4 Anions
- 1.5.5 Other

1.6.0 Adaptation
- 1.6.1 Adaptations, atypical conditions
- 1.6.2 Osmotic adaptation
- 1.6.3 Fe storage

1.7.1 Cell division
2.0.0 Macromolecule metabolism
2.1.0 Macromolecule degradation
- 2.1.1 Degradation of DNA
- 2.1.2 Degradation of RNA
- 2.1.3 Degradation of polysaccharides
- 2.1.4 Degradation of proteins, peptides, glycoproteins

2.2.0 Macromolecule synthesis, modification
- 2.2.01 Amino acyl tRNA synthesis; tRNA modification
- 2.2.02 Basic proteins - synthesis, modification
- 2.2.03 DNA - replication, repair, restriction./modification
- 2.2.04 Glycoprotein
- 2.2.05 Lipopolysaccharide
- 2.2.06 Lipoprotein
- 2.2.07 Phospholipids
- 2.2.08 Polysaccharides - (cytoplasmic)
- 2.2.09 Protein modification
- 2.2.10 Proteins - translation and modification
- 2.2.11 RNA synthesis, modif., DNA transcrip.
- 2.2.12 tRNA

3.0.0 Metabolism of small molecules
3.1.0 Amino acid biosynthesis
- 3.1.01 Alanine
- 3.1.02 Arginine
- 3.1.03 Asparagine
- 3.1.04 Aspartate
- 3.1.05 Chorismate
- 3.1.06 Cysteine
- 3.1.07 Glutamate
- 3.1.08 Glutamine
- 3.1.09 Glycine
- 3.1.10 Histidine
- 3.1.11 Isoleucine
- 3.1.12 Leucine
- 3.1.13 Lysine
- 3.1.14 Methionine
- 3.1.15 Phenylalanine
- 3.1.16 Proline
- 3.1.17 Serine
- 3.1.18 Threonine
- 3.1.19 Tryptophan
- 3.1.20 Tyrosine
- 3.1.21 Valine

3.2.0 Biosynthesis of cofactors, carriers
- 3.2.01 Acyl carrier protein (ACP)
- 3.2.02 Biotin
- 3.2.03 Cobalamin
- 3.2.04 Enterochelin
- 3.2.05 Folic acid
- 3.2.06 Heme, porphyrin
- 3.2.07 Lipoate
- 3.2.08 Menaquinone, ubiquinone
- 3.2.09 Molybdopterin
- 3.2.10 Pantothenate
- 3.2.11 Pyridine nucleotide
- 3.2.12 Pyridoxine
- 3.2.13 Riboflavin
- 3.2.14 Thiamin
- 3.2.15 Thioredoxin, glutaredoxin, glutathione
- 3.2.16 biotin carboxyl carrier protein (BCCP)

Appendix VI (cont):

3.3.0 Central intermediary metabolism
- 3.3.01 2'-Deoxyribonucleotide metabolism
- 3.3.02 Amino sugars
- 3.3.03 Entner-Douderoff
- 3.3.04 Gluconeogenesis
- 3.3.05 Glyoxylate bypass
- 3.3.06 Incorporation metal ions
- 3.3.07 Misc. glucose metabolism
- 3.3.08 Misc. glycerol metabolism
- 3.3.09 Non-oxidative branch, pentose pathway
- 3.3.10 Nucleotide hydrolysis
- 3.3.11 Nucleotide interconversions
- 3.3.12 Oligosaccharides
- 3.3.13 Phosphorus compounds
- 3.3.14 Polyamine biosynthesis
- 3.3.15 Pool, multipurpose conversions of intermed. metabol'm
- 3.3.16 S-adenosyl methionine
- 3.3.17 Salvage of nucleosides and nucleotides
- 3.3.18 Sugar-nucleotide biosynthesis, conversions
- 3.3.19 Sulfur metabolism
- 3.3.20 Amino acids
- 3.3.00 other

3.4.0 Degradation of small molecules
- 3.4.1 Amines
- 3.4.2 Amino acids
- 3.4.3 Carbon compounds
- 3.4.4 Fatty acids
- 3.4.5 Other
- 3.4.0 ATP-proton motive force

3.5.0 Energy metabolism, carbon
- 3.5.1 Aerobic respiration
- 3.5.2 Anaerobic respiration
- 3.5.3 Electron transport
- 3.5.4 Fermentation
- 3.5.5 Glycolysis
- 3.5.6 Oxidative branch, pentose pathway
- 3.5.7 Pyruvate dehydrogenase
- 3.5.8 TCA cycle

3.6.0 Fatty acid biosynthesis
- 3.6.1 Fatty acid and phosphatidic acid biosynthesis

3.7.0 Nucleotide biosynthesis
- 3.7.1 Purine ribonucleotide biosynthesis
- 3.7.2 Pyrimidine ribonucleotide biosynthesis

4.0.0 Cell envelop
- 4.1.0 Periplasmic/exported/lipoproteins
- 4.1.1 Inner membrane
- 4.1.2 Murein sacculus, peptidoglycan
- 4.1.3 Outer membrane constituents
- 4.1.4 Surface polysaccharides & antigens
- 4.1.5 Surface structures

4.2.0 Ribosome constituents
- 4.2.1 Ribosomal and stable RNAs
- 4.2.2 Ribosomal proteins - synthesis, modification
- 4.2.3 Ribosomes - maturation and modification

5.0.0 Extrachromosomal

5.1.0 Laterally acquired elements
- 5.1.1 Colicin-related functions
- 5.1.2 Phage-related functions and prophages
- 5.1.3 Plasmid-related functions
- 5.1.4 Transposon-related functions

6.0.0 Global functions
- 6.1.1 Global regulatory functions

7.0.0 Not classified (included putative assignments)

7.1.1 DNA sites, no gene product

7.2.1 Cryptic genes

Appendix VII: List of colour codes

0 (white) - Pathogenicity/Adaptation/Chaperones
1 (dark grey) - energy metabolism (glycolysis, electron transport etc.)
2 (red) - Information transfer (transcription/translation + DNA/RNA modification)
3 (dark green) - Surface (IM, OM, secreted, surface structures
4 (dark blue) - Stable RNA
5 (Sky blue) - Degradation of large molecules
6 (dark pink) - Degradation of small molecules
7 (yellow) - Central/intermediary/miscellaneous metabolism
8 (light green) - Unknown
9 (light blue) - Regulators
10 (orange) - Conserved hypo
11 (brown) - Pseudogenes and partial genes (remnants)
12 (light pink) - Phage/IS elements
13 (light grey) - Some misc. information e.g. Prosite, but no function

Appendix VIII: List of degenerate nucleotide value/IUB Base Codes.

R = A or G

S = G or C

B = C, G or T

Y = C or T

W = A or T

D = A, G or T

K = G or T

N = A, C, G or T

H = A, C or T

M = A or C

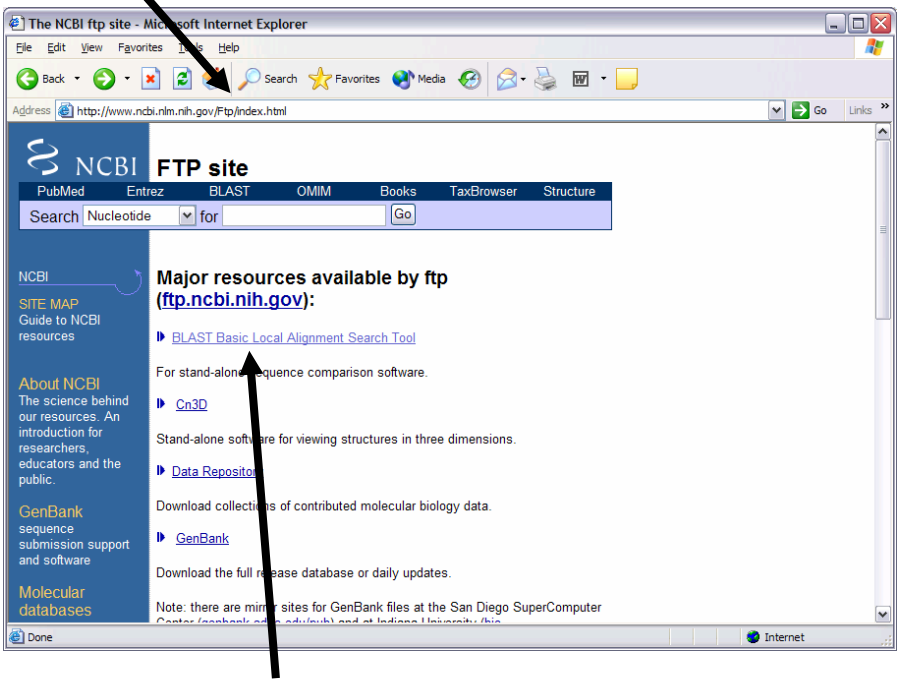
V = A, C or G

Appendix IX: Downloading and installing BLAST on a Windows PC

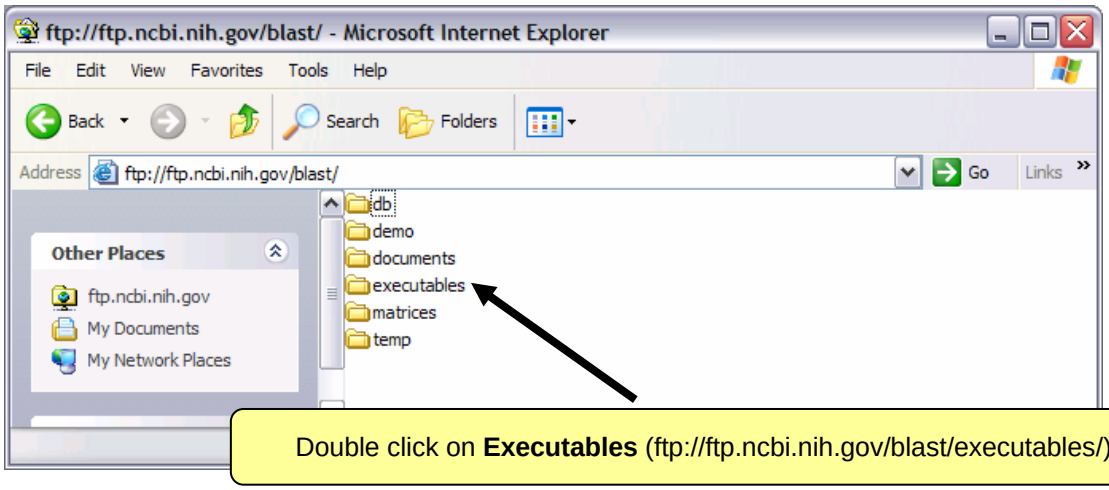
The following pages describe downloading BLAST onto a computer running Windows XP. Downloading onto computers with other versions of Windows should be essentially the same but the windows will look different to the screen shots used here.

Go to NCBI home page (http://www.ncbi.nlm.nih.gov/)

Scroll to bottom, Click on **FTP Site** (left hand side of the screen; http://www.ncbi.nlm.nih.gov/Ftp/index.html)

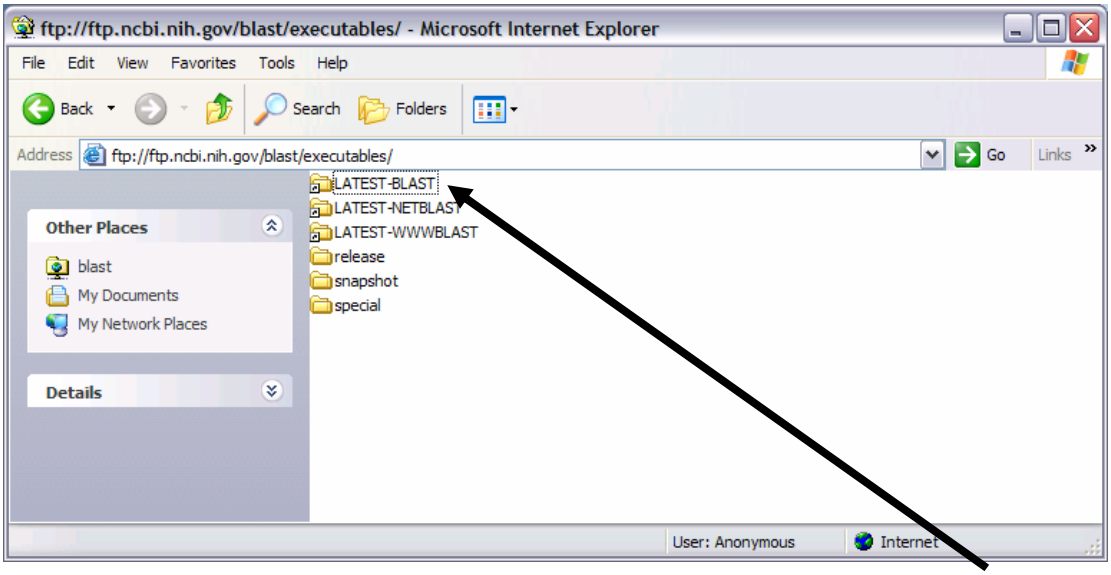


Click on **BLAST Basic Local Alignment Search Tool** (ftp://ftp.ncbi.nih.gov/blast/)

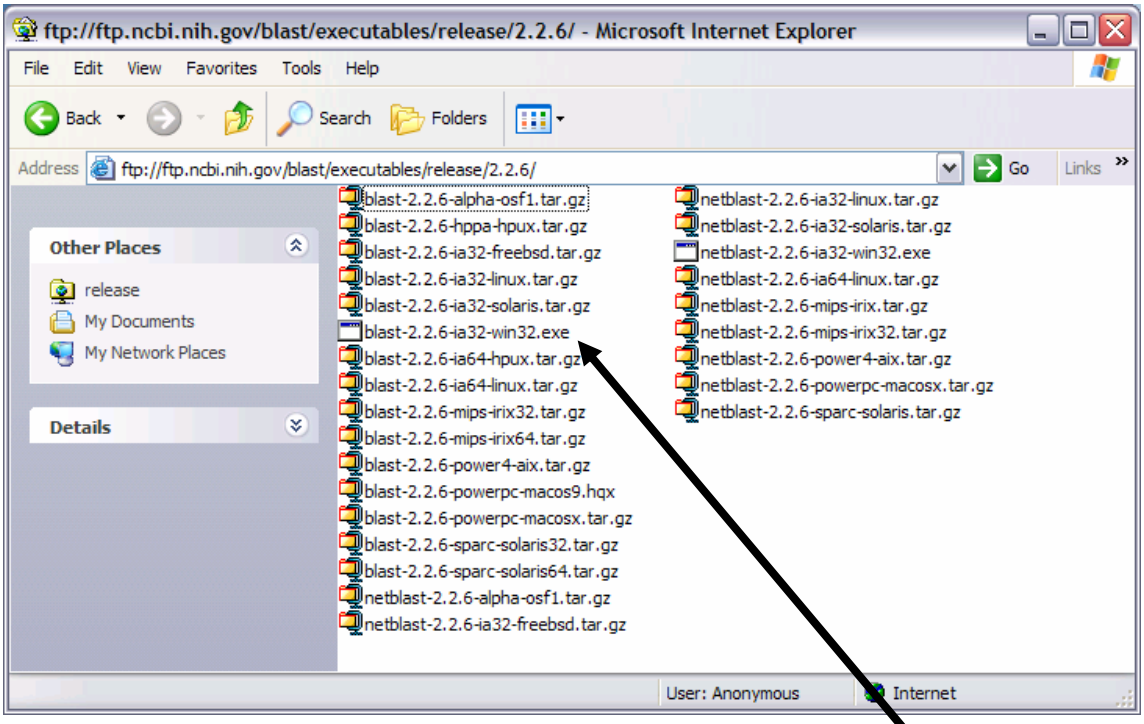


Double click on **Executables** (ftp://ftp.ncbi.nih.gov/blast/executables/)

This page may appear slightly different if you are using Netscape



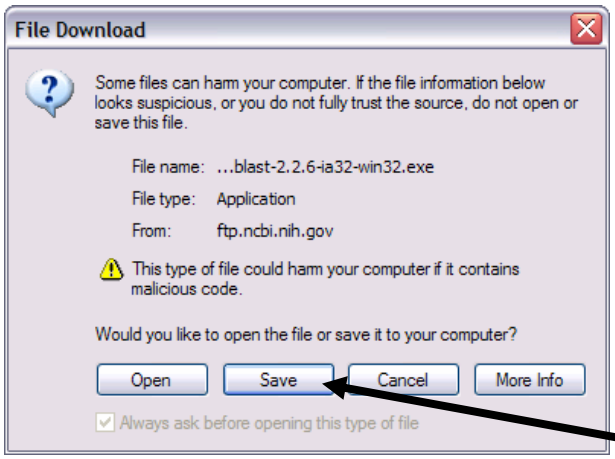
Double click on the **LATEST-BLAST** shortcut



Double click on **blast-2.2.6-ia32-win32.exe**

Blast-2.2.6-ia32win32.exe is the blast exe file for windows

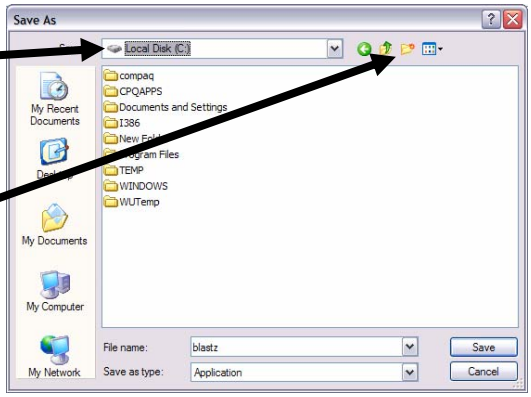
You now need to save the **blast-2.2.6-ia32-win32.exe** file in a new directory, blast, on to the hard drive of your PC

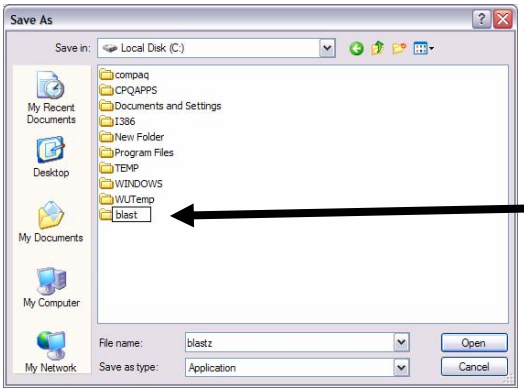


Click on **Save**

Click on **local disk C:**

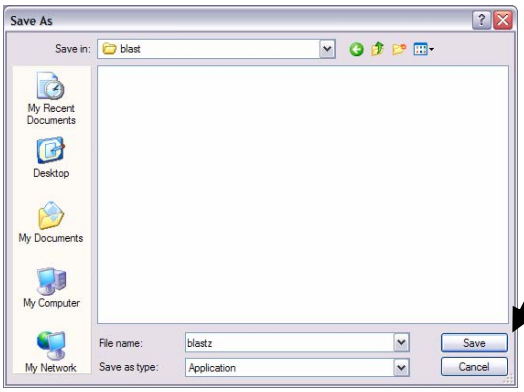
Click on **new directory icon** (folder with a sun peeking through)





Type **blast** in the name box, press **Enter** key.

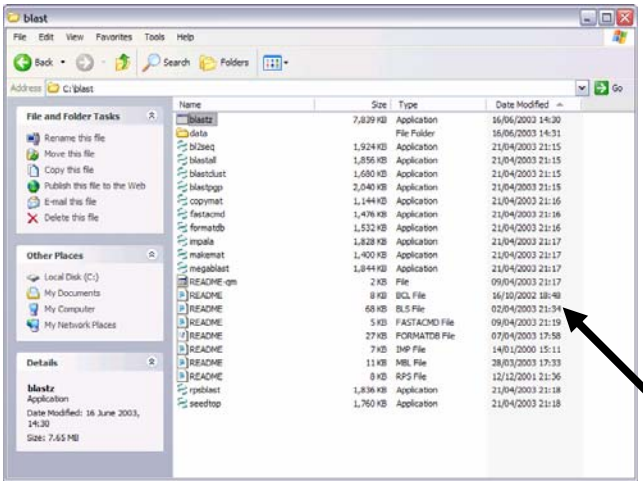
Double click on the new **blast** directory



Click on **Save**

Once downloaded view the contents of the blast directory by clicking on the open folder button

blast-2.2.6-ia32-win32.exe is a compressed file that contains a host of other files.



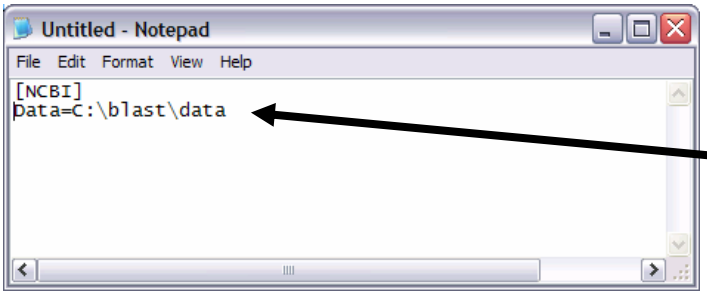
Now double click on the **blast-2.2.6-ia32-win32.exe** file to extract and unpack the rest of the BLAST download files

Included in the directory that has now been unpacked are several README files that describe the various programs in the BLAST software package. These files also provide descriptions of the command line options that you can set when you run the programs. To read these files double click on the icon or view them in notepad.

The **README.BLS** file contains details of the main BLAST program and how to format DNA sequences prior to running BLAST

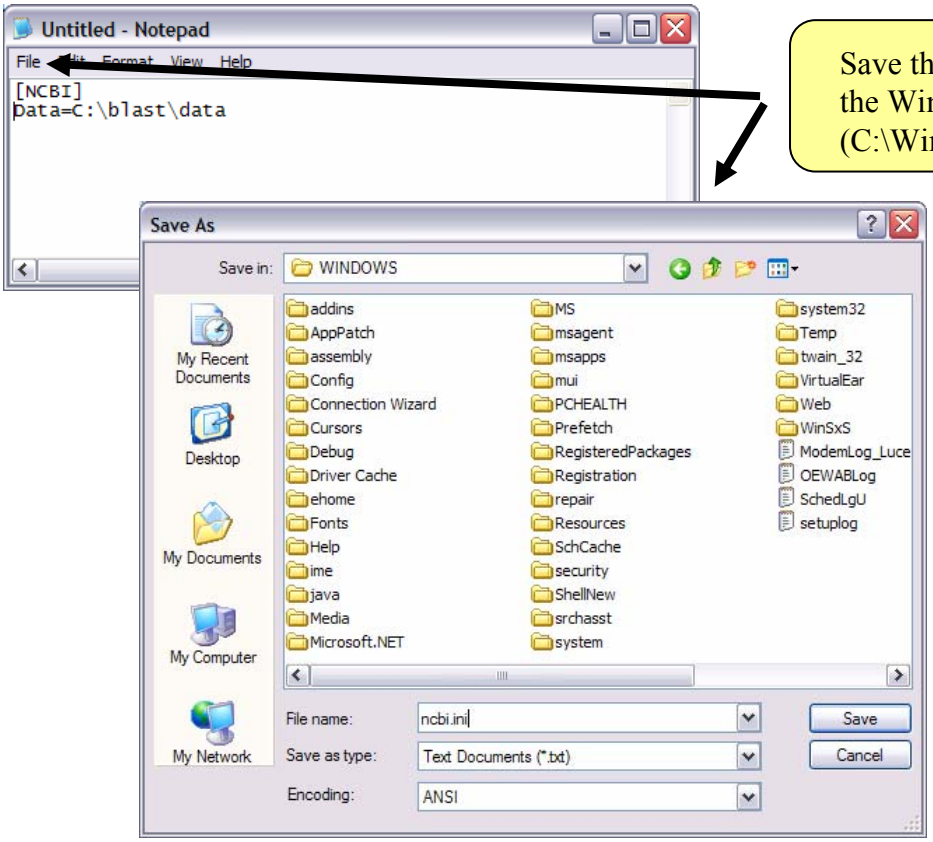
Before you can run BLAST you will need to create an **ncbi.ini** file containing the following lines:

[NCBI]
Data=C:\blast\data



Open **Notepad** (All programs, Accessories menu). Type in the text:

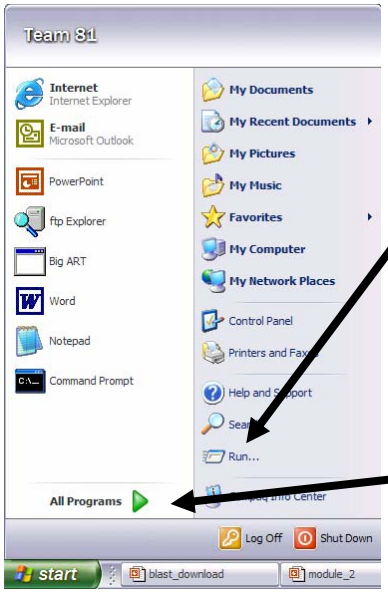
[NCBI]
Data=C:\blast\data



Save the file as **ncbi.ini** in the Windows directory (C:\Windows\)

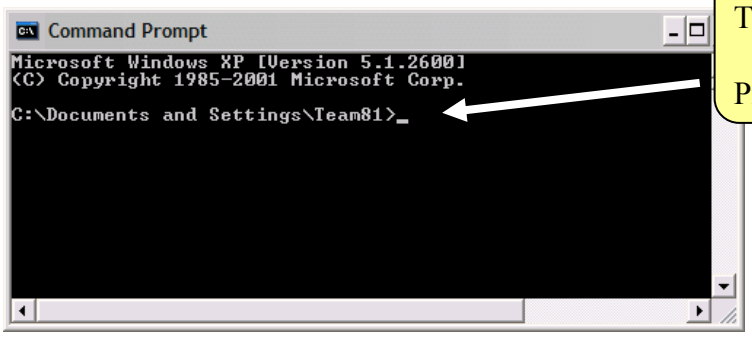
Running BLAST

The BLAST software does not run in Windows, but DOS, an operating system that Windows runs in. When you want to run blast you will need a DOS window a.k.a. Command Prompt



To get to a DOS window select from the Start menu, **Run** and type **Command**

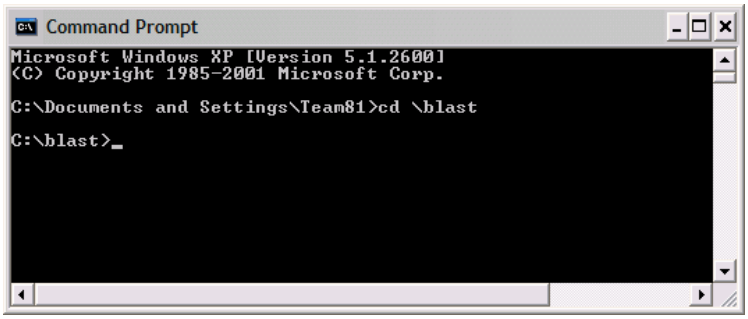
Alternatively the Command prompt icon can be found in Accessories, under All programs



Type **cd \blast**

Press **Return**

This changes the directory to the blast folder which you have just down loaded and unpacked blast-2.2.6-ia32-win32.exe



Now that that you are in the blast directory you can start to run BLAST from the command line

There are several programs in the BLAST package that you have now downloaded that can be used for sequence comparison. For a detailed description of the uses and options see the appropriate README file.